Action at a Distance: A Universal Kernel from Polynomial Alignment and IMQ Distance

Anonymous authors
Paper under double-blind review

Abstract

Inverse-multiquadric (IMQ) kernel sections decay at infinity. We study how quadratic alignment balances their inverse-square decay, adding persistent directional responses while preserving compact-domain approximation. The $\mathtt{E}$ (Yat) kernel multiplies the IMQ kernel by a squared biased inner product. For fixed positive bias and distance offset, its RKHS continuously contains the IMQ RKHS and is universal on every compact domain. This global RKHS inclusion is strict: Yat sections with nonzero centers have a nonzero quadratic directional trace at infinity. Alignment also changes high-dimensional activation covariance. For a fixed number of orthogonal unit centers and fixed kernel parameters, isotropic Gaussian inputs give a rank-one rescaled limiting covariance for IMQ and a full-rank limit for Yat. Under uniform inputs on the radius-$\sqrt{d}$ sphere, both limits are full rank, identifying shared input-norm fluctuations as the Gaussian collapse mechanism in this regime. The connection is constructive: three positive-bias Yat atoms recover any IMQ section exactly, and three are necessary at nonzero centers when the distance offset is shared and biases may vary across atoms. On bounded cubes, rank-r positive-semidefinite quadratics admit r-atom Yat approximations with growing center scales. Accurate IMQ approximation obeys explicit necessary coefficient-variation budgets for $d \geq 4$, without restrictions on width or center locations. Polynomial alignment thus preserves compact-domain approximation while changing global directional behavior and the resources used to represent functions.

1 Introduction

Radial kernel units encode distance to a center, whereas scalar-activation units encode a projection onto a direction. Learned-center RBF and IMQ networks (Broomhead & Lowe, 1988; Moody & Darken, 1989; Park & Sandberg, 1991) use symmetric kernel sections, each constant on spheres about its center. A layer with shared kernel parameters is a finite expansion in one reproducing kernel Hilbert space (RKHS). ReLU and GELU units (Hendrycks & Gimpel, 2016) instead use $\mathbf{w}$ as a direction through $\sigma(\mathbf{w}^\top \mathbf{x} + b)$, but this parameterization does not by itself give a positive-semidefinite kernel on the shared input–parameter domain. We study a kernel section that combines these geometries: it retains an IMQ distance factor while introducing directional dependence through a polynomial alignment factor.

The $\mathtt{E}$ *(Yat) kernel* is

$$k_{\mathtt{E},b,\varepsilon}(\mathbf{w},\mathbf{x}) \;=\; \frac{(\mathbf{w}^\top \mathbf{x} + b)^2}{\|\mathbf{x}-\mathbf{w}\|^2 + \varepsilon}, \qquad b \geq 0,\ \varepsilon > 0. \tag{1}$$

We abbreviate $k_{\mathtt{E},b,\varepsilon}$ to $k_{b,\varepsilon}$. Along $\mathbf{x} = r\mathbf{u}$ with $\|\mathbf{u}\| = 1$, the quadratic numerator term balances the denominator's r^2 growth: the ratio tends to $(\mathbf{u}^\top \mathbf{w})^2$. This degree matching preserves a finite directional response instead of IMQ decay. The kernel is the product of the degree-two polynomial kernel $(\mathbf{w}^\top \mathbf{x} + b)^2$ and the inverse-multiquadric (IMQ) kernel $h_\varepsilon(\mathbf{w},\mathbf{x}) := (\|\mathbf{x}-\mathbf{w}\|^2 + \varepsilon)^{-1}$. The Schur product theorem gives positive semidefiniteness

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

directly (Schölkopf & Smola, 2002; Steinwart & Christmann, 2008). Our focus is the induced function space: which IMQ guarantees survive the product, and which functions become available through alignment.

The numerator makes this relationship explicit. Its expansion separates radial, linear-alignment, and quadratic-alignment channels. Varying the bias also allows the alignment terms to cancel: for every $t > 0$,

$$k_{3t,\varepsilon}(\mathbf{w}, \cdot) - 2k_{2t,\varepsilon}(\mathbf{w}, \cdot) + k_{t,\varepsilon}(\mathbf{w}, \cdot) = 2t^2 h_\varepsilon(\mathbf{w}, \cdot). \tag{2}$$

This second finite difference recovers the IMQ section using three distinct positive biases and one shared distance offset. It acts across an atom family, rather than within a single fixed-bias dictionary.

Contributions. We establish inherited IMQ guarantees, then characterize the kernel-specific directional, covariance, and exact-representation consequences of alignment.

1. *Fixed-kernel inheritance and its zero-bias boundary.* For fixed $b > 0$ and $\varepsilon > 0$, $b^2 h_\varepsilon \preceq k_{b,\varepsilon}$ in the Loewner order, so the IMQ RKHS embeds continuously into $\mathcal{H}_{b,\varepsilon}$. This gives universality, injective finite-signed-measure embeddings, and strict positive definiteness on every compact domain; the domination constant b^2 is globally sharp. At $b = 0$, every RKHS function vanishes at the origin; the kernel is universal on a nonempty compact set if and only if that set excludes the origin (Section 2).
2. *Directional trace and strict global enlargement.* Along a ray $r\mathbf{u}$ with $\|\mathbf{u}\| = 1$, a Yat section converges to $(\mathbf{u}^\top \mathbf{w})^2$, while every function in the global IMQ RKHS, and hence every finite IMQ expansion, vanishes at infinity. The resulting quadratic trace distinguishes the finite spans and witnesses strict global RKHS containment for $b > 0$. Compact-domain universality therefore coexists with different behavior at infinity (Section 3).
3. *Input fluctuations and center geometry in activation covariance.* With a fixed number of unit centers and fixed kernel parameters, we compare Gaussian and fixed-norm inputs as dimension grows. For orthogonal centers, the leading rescaled IMQ covariance is rank one under isotropic Gaussian inputs but full rank on the radius-$\sqrt{d}$ sphere; the Yat limit is full rank in both cases. The general-center limits distinguish shared input-norm fluctuations from the linear and quadratic geometry of the centers (Section 3).
4. *Exact recovery and resource-constrained approximation.* The three-atom identity (2) is optimal for exact recovery on all of $\mathbb{R}^d$ at nonzero centers when the distance offset is shared; at the origin, one atom suffices. A single scaled atom can instead approximate an IMQ section uniformly on all of $\mathbb{R}^d$ as its bias grows without bound.

   For resource-constrained approximation of quadratic targets on bounded cubes, rank-r positive-semidefinite forms admit r-atom Yat approximations with growing center scales, whereas accurate IMQ approximation requires a trace-sensitive coefficient budget for $d \geq 4$ and an additional eigenvector-slice bound for $d \geq 5$ (Section 4).

As an application, shared-(b, ε) layers admit the standard exact Gram-norm identity. The Yat kernel diagonal then gives an explicit Rademacher bound for a fixed RKHS norm budget (Section 5).

Fixing a network prefix extends this geometry to a pullback RKHS with an exact quotient-norm characterization. Useful parameter-independent capacity control across depth remains open.

2 What Distance Retains: Fixed-Kernel Universality

Positive bias retains an IMQ component within the Yat kernel. The remaining alignment components are positive semidefinite, so this decomposition yields kernel-order domination and transfers the IMQ approximation guarantees.

Notation. Unless stated otherwise, $d \geq 1$, $\mathcal{X} \subset \mathbb{R}^d$ is compact, $b \geq 0$, and $\varepsilon > 0$. Write $D_\varepsilon(\mathbf{x}, \mathbf{w}) := \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$ for the offset squared distance, and set

$$g_\varepsilon(\mathbf{x}; \mathbf{w}, b) := \frac{(\mathbf{x}^\top \mathbf{w} + b)^2}{D_\varepsilon(\mathbf{x}, \mathbf{w})} \qquad \text{(biased Yat atom)},$$
$$h_\varepsilon(\mathbf{x}, \mathbf{w}) := k^\varepsilon_{\mathrm{IMQ}}(\mathbf{x}, \mathbf{w}) := D_\varepsilon(\mathbf{x}, \mathbf{w})^{-1} \quad \text{(IMQ)}.$$

The notation $g_\varepsilon(\mathbf{x}; \mathbf{w}, b)$ emphasizes the atom's parameters; fixing (b, ε) gives the symmetric kernel $k_{b,\varepsilon}(\mathbf{w}, \mathbf{x})$. The unbiased section is $k_{\mathbb{E}}(\mathbf{w}, \mathbf{x}) = g_\varepsilon(\mathbf{x}; \mathbf{w}, 0)$.

From channels to inherited guarantees. Expanding the numerator gives

$$k_{b,\varepsilon} = \underbrace{b^2 h_\varepsilon}_{\text{radial}} + \underbrace{2b(\mathbf{x}^\top \mathbf{w}) h_\varepsilon}_{\text{linear alignment}} + \underbrace{(\mathbf{x}^\top \mathbf{w})^2 h_\varepsilon}_{\text{quadratic alignment}} .$$

Each channel is PSD by the Schur product theorem. In particular, $k_{b,\varepsilon} - b^2 h_\varepsilon$ is PSD, which is the meaning of $b^2 h_\varepsilon \preceq k_{b,\varepsilon}$ in the Loewner order. For positive bias, kernel-order domination embeds the IMQ RKHS and transfers universality.

For $b > 0$, the sharp kernel order gives the continuous embedding

$$\mathcal{H}_{h_\varepsilon} \subseteq \mathcal{H}_{b,\varepsilon}, \qquad \|f\|_{\mathcal{H}_{b,\varepsilon}} \leq b^{-1} \|f\|_{\mathcal{H}_{h_\varepsilon}}.$$

On every compact domain, the fixed kernel is universal, its mean embedding separates finite signed Borel measures, its MMD metrizes weak convergence of probability measures, and it is strictly positive definite. At $b = 0$, universality on a nonempty compact $\mathcal{X}$ holds exactly when $\mathbf{0} \notin \mathcal{X}$. Proposition 2 collects the complete statement.

Appendix B gives the individual statements and proofs. The constant b^2 is sharp on $\mathbb{R}^d$; the associated RKHS norm bound scales as $1/b$. Thus positive bias preserves the IMQ space, but the transferred norm bound deteriorates as $b \downarrow 0$.

The channel decomposition also describes the RKHS itself: it is the sum of the three channel spaces, with squared norm given by the minimum sum of squared component norms (Theorem 8). We take the RKHS of the identically-zero kernel to be $\{0\}$ with trivial norm, so this description includes the endpoint $b = 0$, where only the quadratic channel remains. At that endpoint every RKHS function vanishes at the origin, explaining the obstruction on compact domains that contain it.

Fixed kernel versus atom family. These guarantees concern one shared pair (b, ε). Allowing parameters to vary across atoms instead gives the broader finite span $F^{\geq 0}_{\mathbb{E}} := \mathrm{span}\{g_\eta(\cdot; \mathbf{w}, \beta) : \mathbf{w} \in \mathbb{R}^d,\ \beta \geq 0,\ \eta > 0\}$, not a single shared-parameter kernel expansion. Unlike IMQ and Gaussian RBF sections, nonzero-center Yat sections retain a directional trace; unlike polynomial sections, each is bounded. Appendix A.3, Table 2, gives the full comparison.

3 What Alignment Adds: Nonradial Structure

The IMQ embedding leaves two questions: which global functions alignment adds, and how it changes activation geometry under a specified input distribution. We first use directional limits to show that the embedding is strict. We then compare Gaussian and fixed-norm activation covariances, and identify the quadratic center geometry that survives input normalization.

3.1 Directional traces and strict RKHS enlargement

The radial baseline is a decay property of the entire IMQ RKHS, not just its finite expansions.

Lemma 1 (IMQ RKHS functions vanish at infinity)**.** *Every $f \in \mathcal{H}_{h_\varepsilon}$ belongs to $C_0(\mathbb{R}^d)$, i.e., $f(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$.*

The lemma is a *global* statement on $\mathbb{R}^d$, distinct from compact-domain universality: the IMQ RKHS restricts densely to $C(\mathcal{X})$ on every compact $\mathcal{X}$ (Micchelli et al., 2006), yet its ambient functions vanish at infinity.

The quadratic channel carries the directional trace. Write $F_{\mathrm{IMQ},\varepsilon} := \mathrm{span}\{h_\varepsilon(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d\}$ for the finite IMQ span. The following proposition identifies the trace range of the Yat span with the homogeneous quadratic forms on $\mathbb{S}^{d-1}$.

Proposition 1 (Directional asymptotic trace separates Yat from IMQ)**.** *Let $\mathcal{D}_{T_\infty}$ denote the linear subspace of continuous functions $F : \mathbb{R}^d \to \mathbb{R}$ for which the radial limit $\lim_{r\to\infty} F(r\mathbf{u})$ exists for every $\mathbf{u} \in \mathbb{S}^{d-1}$, and define on $\mathcal{D}_{T_\infty}$ the asymptotic trace operator $T_\infty F(\mathbf{u}) := \lim_{r\to\infty} F(r\mathbf{u})$. Both $F_{\mathbb{E}}^{\geq 0}$ and $F_{\mathrm{IMQ},\varepsilon}$ lie in $\mathcal{D}_{T_\infty}$.*

(a) Atom trace. *For every $\mathbf{w} \in \mathbb{R}^d$, every $b \in \mathbb{R}$, and every $\mathbf{u} \in \mathbb{S}^{d-1}$, the biased Yat atom satisfies*

$$T_\infty\, g_\varepsilon(\cdot; \mathbf{w}, b)(\mathbf{u}) \;=\; (\mathbf{u}^\top \mathbf{w})^2, \tag{3}$$

independent of b and ε, and the convergence is uniform in $\mathbf{u}$.

(b) IMQ null trace. *$T_\infty F \equiv 0$ for every finite IMQ combination $F \in F_{\mathrm{IMQ},\varepsilon}$.*

(c) Trace range. *The induced linear map $T_\infty : F_{\mathbb{E}}^{\geq 0} \to C(\mathbb{S}^{d-1})$ is surjective onto the space of restrictions of homogeneous quadratic forms, $\mathcal{Q}_2(\mathbb{S}^{d-1}) = \{\mathbf{u} \mapsto \mathbf{u}^\top \mathbf{A}\mathbf{u} : \mathbf{A} \in \mathrm{Sym}(d)\}$, of dimension $d(d{+}1)/2$, while $F_{\mathrm{IMQ},\varepsilon} \subseteq \ker T_\infty$.*

A Yat section with a nonzero center has a nonzero directional trace and hence cannot belong to the IMQ RKHS. Together with the inclusion in Proposition 2, this gives the following global separation.

Corollary 1 (Strict containment and non-reversibility)**.** *For $b > 0$, $\varepsilon > 0$: $\mathcal{H}_{h_\varepsilon} \subsetneq \mathcal{H}_{b,\varepsilon}$. Moreover, no constant $C > 0$ satisfies $k_{b,\varepsilon} \preceq C\, h_\varepsilon$ on $\mathbb{R}^d$.*

Small RKHS norm need not imply a small far-field trace: Proposition 6 constructs finite expansions tending to zero in RKHS norm with a fixed nonzero trace. Thus the finite-span trace has no continuous extension to the ambient RKHS.

3.2 Gaussian versus fixed-norm activation covariance

The trace distinguishes global function spaces. Activation covariance asks a different question: how do responses vary together under a given input law? For isotropic Gaussian inputs and a fixed number of unit centers, the shared fluctuation of $\|\mathbf{x}\|^2$ produces a rank-one leading IMQ covariance. Fixing the input norm removes this fluctuation. The Yat numerator retains directional variation through squared biased projections, leading to the linear and quadratic covariance terms below.

Theorem 1 (Activation covariance under Gaussian and fixed-norm inputs)**.** *Fix $m \in \mathbb{N}$, $b \geq 0$, and $\varepsilon > 0$. For each $d \geq m$, let $\mathbf{w}_{1,d}, \ldots, \mathbf{w}_{m,d} \in \mathbb{R}^d$ be unit vectors whose Gram matrices $\mathbf{C}_d = [\mathbf{w}_{i,d}^\top \mathbf{w}_{j,d}]$ converge to $\mathbf{C} \in \mathbb{R}^{m\times m}$. For either input law below, define*

$$\mathbf{H}_d := \big(h_\varepsilon(\mathbf{x}_d, \mathbf{w}_{j,d})\big)_{j=1}^m, \qquad \mathbf{Y}_d := \big(k_{b,\varepsilon}(\mathbf{w}_{j,d}, \mathbf{x}_d)\big)_{j=1}^m.$$

Both activation vectors are columns. The following limits hold entrywise and in operator norm, where $\circ$ denotes the Hadamard product:

(i) For $\mathbf{x}_d \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$,

$$d^3\,\mathrm{Cov}(\mathbf{H}_d) \longrightarrow 2\mathbf{1}\mathbf{1}^\top, \qquad d^2\,\mathrm{Cov}(\mathbf{Y}_d) \longrightarrow 2(\mathbf{C}\circ\mathbf{C}) + 4b^2\mathbf{C}. \tag{4}$$

(ii) For $\mathbf{x}_d$ uniform on $\sqrt{d}\,\mathbb{S}^{d-1}$,

$$d^4\,\mathrm{Cov}(\mathbf{H}_d) \longrightarrow 4\mathbf{C}, \qquad d^2\,\mathrm{Cov}(\mathbf{Y}_d) \longrightarrow 2(\mathbf{C}\circ\mathbf{C}) + 4b^2\mathbf{C}. \tag{5}$$

For orthonormal centers, the Gaussian IMQ limit has eigenvalues $(2m, 0, \ldots, 0)$ and the spherical IMQ limit is $4\mathbf{I}_m$. The Yat limit is $(2 + 4b^2)\mathbf{I}_m$ in both cases. Thus the limiting effective ranks are 1 versus m under Gaussian input and m for both kernels under spherical input.

The unscaled covariances vanish with d in both models; the limits describe their leading nonzero terms. Both proofs appear in Appendix C.2.

Finite-dimensional scope. For $d \geq 34$, Proposition 9 gives explicit finite-dimensional operator-norm errors: $O_{b,\varepsilon}(m/\sqrt{d})$ in the Gaussian model and $O_{b,\varepsilon}(m/d)$ on the sphere, relative to the actual center Gram matrix $\mathbf{C}_d$. Corollary 5 additionally tracks $\|\mathbf{C}_d - \mathbf{C}\|_{\mathrm{op}}$ and the limiting spectral gap. Thus these errors vanish for $m = o(\sqrt{d})$ and $m = o(d)$, respectively. These are population bounds, not joint sample-spectrum or trained-prefix guarantees.

From population covariance to sample spectra. For a nonzero PSD matrix, define the participation-ratio effective rank by $r_{\mathrm{eff}}(\mathbf{A}) := \mathrm{tr}(\mathbf{A})^2/\mathrm{tr}(\mathbf{A}^2)$. For n i.i.d. inputs from the specified distribution, the empirical centered activation covariance converges almost surely to its population counterpart as $n \to \infty$. Taking first $n \to \infty$ and then $d \to \infty$, the effective rank under Gaussian input converges to 1 for IMQ and to m for orthogonal Yat centers; for orthogonal centers under fixed-norm spherical input it converges to m for both. These sequential limits use fixed m; the separate bounds in Proposition 9 quantify population errors for varying m, but do not handle arbitrary data-dependent center selection.

If $\mathbf{A} \in \mathbb{R}^{n\times m}$ is the activation matrix and $\mathbf{P}_n = \mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n$, the centered sample Gram $\mathbf{P}_n\mathbf{A}\mathbf{A}^\top\mathbf{P}_n/n$ and activation covariance $\mathbf{A}^\top\mathbf{P}_n\mathbf{A}/n$ have the same nonzero eigenvalues. The theorem therefore describes the leading centered Gram spectrum as well. Figure 1 visualizes the Gaussian and fixed-norm limits with exact sampling; Appendix F.1 gives the diagnostic protocol.

3.3 Quadratic lifting after normalization

Removing input-norm fluctuations leaves a distinction between linear and quadratic center geometry. The spherical IMQ limit is $4\mathbf{C}$, so its rank is the dimension spanned by the limiting center coordinates. The Yat limit additionally contains $2(\mathbf{C}\circ\mathbf{C})$, the Gram matrix of their quadratic features. These features can distinguish activation directions that are linearly dependent in the original center coordinates.

For a concrete example, take three centers in a fixed two-dimensional subspace, embedded in $\mathbb{R}^d$ as d grows:

$$\mathbf{v}_1 = \mathbf{e}_1, \qquad \mathbf{v}_2 = \mathbf{e}_2, \qquad \mathbf{v}_3 = (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}.$$

Here $\mathrm{rank}\,\mathbf{C} = 2$, but $\det(\mathbf{C}\circ\mathbf{C}) = 1/2 > 0$. At $b = 0$, the spherical IMQ covariance limit therefore has rank two and the Yat limit has rank three (Figure 2). The rank-three conclusion also holds for $b > 0$. Normalization removes shared radial fluctuations; polynomial alignment can still add independent covariance directions. These rank statements concern

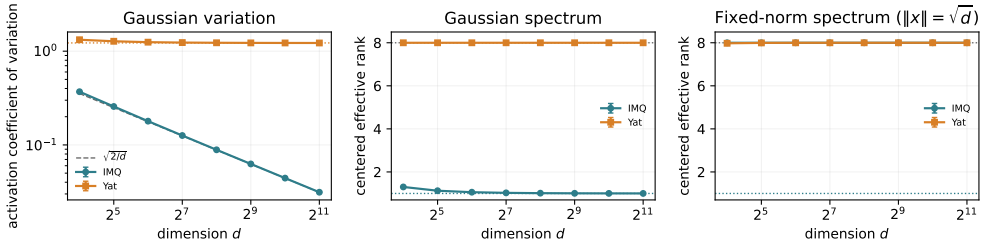

Figure 1: **Gaussian and fixed-norm inputs give different leading IMQ covariances.** Mean ± standard deviation over five Monte Carlo seeds, with 10^5 isotropic Gaussian or fixed-norm spherical inputs per seed, $m = 8$ orthogonal unit centers, and $b = \varepsilon = 1$. Under Gaussian input, IMQ follows the predicted $\sqrt{2/d}$ coefficient of variation and its leading centered covariance approaches rank one, while Yat retains effective rank 8. On the sphere $\|\mathbf{x}\| = \sqrt{d}$, both effective ranks approach 8, matching Theorem 1.

the leading rescaled covariance limits under the theorem's fixed-number-of-centers asymptotics. The following corollary gives the corresponding criterion for general limiting center geometry.

Corollary 2 (Rank criterion for general limiting center geometry)**.** *Under the assumptions of Theorem 1, factor* $\mathbf{C} = \mathbf{V}^\top \mathbf{V}$ *with columns* $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathbb{R}^r$*, where* $r = \operatorname{rank} \mathbf{C}$*. The common rescaled Yat limit is the Gram matrix of the lifted vectors*

$$\boldsymbol{\psi}_b(\mathbf{v}_j) = \big(\sqrt{2}\,\mathbf{v}_j \otimes \mathbf{v}_j,\ 2b\,\mathbf{v}_j\big).$$

Consequently, its rank equals $\dim \operatorname{span}\{\boldsymbol{\psi}_b(\mathbf{v}_j) : 1 \le j \le m\}$*, whereas the rescaled spherical IMQ limit has rank* r*. The Yat limit is full rank exactly when the lifted vectors are linearly independent. Its rank is at most* $\min\{m, r(r+1)/2\}$ *for* $b = 0$*, and at most* $\min\{m, r(r+1)/2 + r\}$ *for* $b > 0$*.*

The proof is in Appendix C.2.

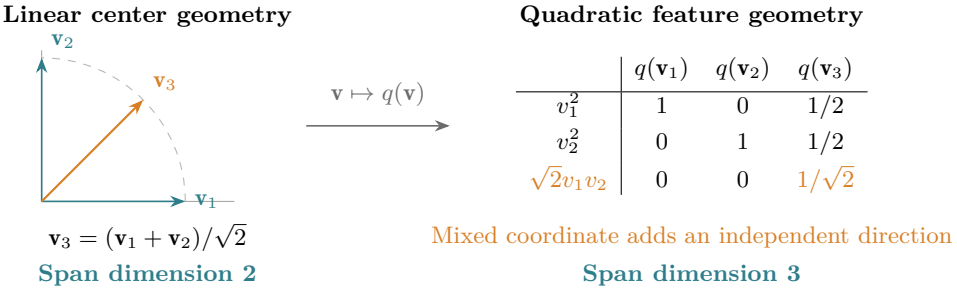

Figure 2: **Quadratic alignment separates linearly dependent centers.** The three unit centers span two dimensions. Their quadratic coordinates $q(\mathbf{v}) = (v_1^2, v_2^2, \sqrt{2}v_1v_2)$ span three: only the third center has a nonzero mixed coordinate. Since $q(\mathbf{v}_i)^\top q(\mathbf{v}_j) = (\mathbf{v}_i^\top \mathbf{v}_j)^2$, this gives spherical covariance-limit ranks two for IMQ and three for Yat, already at $b = 0$.

4 Constructive Representation and Approximation

Fixed positive bias gives continuous RKHS containment, but containment alone does not specify a finite representation. We first show how varying the bias recovers IMQ sections exactly. We then distinguish exact recovery from uniform approximation and compare the resources used to approximate quadratic targets.

4.1 Exact IMQ recovery and approximation transfer

Bias finite differences recover IMQ sections exactly; the next theorem also establishes the sharp atom count at nonzero centers.

Theorem 2 (Exact IMQ recovery with sharp three-atom complexity)**.** *Fix* $d \ge 1$ *and* $\varepsilon > 0$*.*

(i) For every $\mathbf{w} \in \mathbb{R}^d$ *and* $h > 0$*, the positive-bias atoms satisfy*

$$g_\varepsilon(\mathbf{x}; \mathbf{w}, 3h) \;-\; 2\, g_\varepsilon(\mathbf{x}; \mathbf{w}, 2h) \;+\; g_\varepsilon(\mathbf{x}; \mathbf{w}, h) \;=\; \frac{2h^2}{\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon} \;=\; 2h^2\, k^\varepsilon_{\mathrm{IMQ}}(\mathbf{x}, \mathbf{w}), \quad (6)$$

for every $\mathbf{x} \in \mathbb{R}^d$*. Consequently* $F_{\mathrm{IMQ},\varepsilon} \subseteq F^+_{\mathbb{E},\varepsilon}$*, where* $F^+_{\mathbb{E},\varepsilon} = \operatorname{span}\{g_\varepsilon(\cdot; \mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d, b > 0\}$ *and* $F_{\mathrm{IMQ},\varepsilon} = \operatorname{span}\{k^\varepsilon_{\mathrm{IMQ}}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d\}$*. The containment is strict.*

(ii) For every $\mathbf{w}_0 \neq \mathbf{0}$*, no linear combination of at most two biased Yat atoms, all using this same* ε *but with arbitrary centers and real biases, equals a nonzero scalar multiple of* $k^\varepsilon_{\mathrm{IMQ}}(\cdot, \mathbf{w}_0)$*.*

At $\mathbf{w}_0 = \mathbf{0}$*, one atom suffices:* $g_\varepsilon(\cdot; \mathbf{0}, b)/b^2 = h_\varepsilon(\cdot, \mathbf{0})$ *for* $b > 0$*.*

The finite-difference identity is elementary; minimality requires a separate argument. Appendix D.1 proves minimality using irreducible quadrics for $d \geq 2$ and pole matching for $d = 1$, and gives the full recovery proof.

Exact equality versus approximation with growing bias. The three-atom lower bound concerns exact equality. A single scaled atom can instead approximate an IMQ section uniformly on the entire input space: for fixed $\mathbf{w}$ and $\varepsilon > 0$,

$$\left\| B^{-2} g_\varepsilon(\cdot; \mathbf{w}, B) - h_\varepsilon(\cdot, \mathbf{w}) \right\|_{L^\infty(\mathbb{R}^d)} = O_{\mathbf{w},\varepsilon}(B^{-1}) \longrightarrow 0.$$

Lemma 5 gives an explicit bound. This uses unbounded bias and a varying kernel, not exact equality or approximation within one fixed-bias dictionary. Atom count alone does not describe the full cost.

Approximation transfer. For fixed $\varepsilon > 0$ and $h > 0$, every m-atom IMQ expansion has an identical Yat representation with at most $3m$ atoms and biases in $\{h, 2h, 3h\}$. Thus an IMQ error rate $\phi(m)$ transfers to $\phi(\lfloor M/3 \rfloor)$ at Yat budget $M \geq 3$. This also gives compact-domain universality of the variable-bias span. Corollaries 7 and 8 in Appendix D.2 state these consequences in full.

4.2 Quadratic approximation and coefficient budgets

Exact recovery transfers IMQ representations to Yat. To compare approximation resources beyond that transfer, we consider quadratic targets directly. The Yat construction controls atom count while allowing center scales to grow; the IMQ obstruction instead controls total coefficient variation, without restricting width or center locations. For fixed $\varepsilon > 0$, define

$$\mathcal{V}_{\mathrm{IMQ}}(\Lambda) := \left\{ \textstyle\sum_{j=1}^m a_j h_\varepsilon(\cdot, \mathbf{v}_j) : m \in \mathbb{N},\ \sum_j |a_j| \leq \Lambda,\ \mathbf{v}_j \in \mathbb{R}^d \right\}.$$

Theorem 3 (Quadratic approximation and coefficient-variation lower bounds)**.** *Let $\mathbf{A} \succeq 0$ have rank r and largest eigenvalue $\lambda_{\max} > 0$. Fix $R > 0$, $\varepsilon > 0$, and $\delta > 0$.*

(i) There is an unbiased Yat expansion with at most r atoms that approximates $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ uniformly on $[-R, R]^d$ to error δ.

(ii) Suppose $d \geq 5$. If $F \in \mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ attains the same error, then necessarily

$$\Lambda \geq \sup_{0 \leq t \leq 1} \left(\lambda_{\max} R^2 t - \delta\right)_+ \left(R^2(1-t) + \varepsilon\right). \tag{7}$$

In particular, taking $t = 1/4$ gives

$$\Lambda \geq \left(\frac{\lambda_{\max} R^2}{4} - \delta \right)_+ \left(\frac{3R^2}{4} + \varepsilon \right).$$

(iii) Already for $d \geq 4$, every such IMQ approximant satisfies

$$\Lambda \geq \left(\frac{R^2 \operatorname{tr} \mathbf{A}}{d} - \delta \right)_+ (R^2 + \varepsilon). \tag{8}$$

For $d \geq 5$, both bounds apply, so their maximum is a valid lower bound.

Resource interpretation. Parts (ii)–(iii) give necessary IMQ coefficient budgets for a prescribed error, regardless of the number or location of centers. These lower bounds remain finite as $\delta \downarrow 0$; they do not establish a coefficient-divergence rate. Part (i) gives a Yat atom-count upper bound, achieved by increasing center scale as the target error decreases; the construction also covers arbitrary symmetric quadratics (Proposition 10). For fixed nonzero target, d, R, ε, the prescribed scales give an RKHS norm of order δ^{-2} as $\delta \downarrow 0$ (Remark 4). The comparison is therefore between different resources: it supplies neither a matching atom-count lower bound nor an improved generalization guarantee through the layer-local Rademacher bound.

Appendix D.3 proves the construction and both coefficient-budget bounds, with a worked example and mixed-target extension; Appendix C.3.1 gives the exterior-shell bounds.

5 Layer-Local RKHS Consequences

Standard RKHS identities give exact layer norms and fixed-prefix geometry; the Yat diagonal makes their capacity consequences explicit. These applications do not provide capacity control over learned prefixes.

Ambient layer geometry. For a layer with shared $b \geq 0$ and $\varepsilon > 0$, write $f = \sum_j \alpha_j k_{b,\varepsilon}(\mathbf{w}_j, \cdot)$ and $\mathbf{K}_{ij} = k_{b,\varepsilon}(\mathbf{w}_i, \mathbf{w}_j)$. Its global RKHS norm is

$$\|f\|^2_{\mathcal{H}_{b,\varepsilon}} = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha},$$

including repeated centers (Proposition 13). This quantity is computable from trained centers and readout coefficients. For multiclass logits, centering the class coefficients removes common-logit shifts (Corollary 9). The shared-kernel interpretation requires shared bias and distance offset; per-neuron biases instead select atoms from the variable-bias family.

For n inputs of Euclidean norm at most R and fixed (B, b, ε), the kernel diagonal bounds the empirical Rademacher complexity of a radius-B RKHS ball by $B(R^2+b)/\sqrt{n\varepsilon}$. Appendix E gives the precise sample assumptions and data-dependent refinement. If $0 \leq b \leq b_{\max}$ and $\varepsilon \geq \varepsilon_{\min} > 0$, Proposition 14 embeds these balls contractively into $\mathcal{H}_{b_{\max},\varepsilon_{\min}}$. The envelope and norm budget must be fixed independently; a post-hoc fitted norm alone gives no uniform guarantee.

Geometry after a fixed prefix. Let Φ be a fixed continuous prefix on a compact input domain $\mathcal{X}$. The next layer induces the pullback kernel

$$\widetilde{k}(\mathbf{x}, \mathbf{x}') = k_{b,\varepsilon}\big(\Phi(\mathbf{x}), \Phi(\mathbf{x}')\big).$$

Its RKHS contains the composed function $f \circ \Phi$. The exact pullback norm is the minimum ambient RKHS norm among functions agreeing with f on $\Phi(\mathcal{X})$; it can therefore be smaller than the computable ambient Gram norm. Restriction to an input domain is the same minimum-extension construction with the inclusion map as prefix (Appendix E.3).

For $b > 0$, the pullback kernel is universal on $\mathcal{X}$ exactly when Φ is injective. If a prefix identifies two inputs, every pullback-RKHS function agrees at those inputs, preventing universality. These statements describe the RKHS conditional on a fixed prefix, not a parameter-independent capacity bound for the complete trained network.

Diagnostics. Appendix F reports covariance illustrations, descriptive frozen-CLIP norm diagnostics, and 261M language-model feasibility runs at retained settings. These are not matched-tuning comparisons and establish neither a generalization advantage nor a global depth-capacity guarantee.

6 Conclusion and Discussion

Compact-domain universality does not determine global behavior or representation cost. The Yat construction makes this distinction explicit: for fixed positive bias and distance offset it contains the IMQ RKHS, yet its nonzero directional traces extend beyond the decaying radial space. The three-atom finite-difference identity connects the two families constructively without making their finite representations equivalent. Its optimality for exact recovery at nonzero centers with a shared distance offset, together with one-atom global uniform approximation at growing bias, separates exact equality from approximation with unbounded parameters.

The covariance analysis identifies a different role for alignment. With a fixed number of orthogonal unit centers and fixed kernel parameters, Gaussian inputs give a rank-one leading rescaled IMQ covariance, whereas the Yat limit is full rank. Fixing the input norm removes this IMQ collapse: both limits are then full rank for orthogonal centers. For general center geometry, the quadratic lift $\mathbf{C} \circ \mathbf{C}$ can still add covariance directions beyond those in $\mathbf{C}$. These results distinguish shared radial fluctuations from alignment geometry rather than asserting an unconditional deficiency of radial kernels.

Representation cost likewise depends on the resource being controlled. The quadratic constructions achieve small Yat atom count through growing center scales and RKHS norms; the IMQ lower bounds instead constrain coefficient variation while leaving width and centers unrestricted. These are approximation tradeoffs, not a generalization advantage or a coefficient-divergence rate. The numerical diagnostics illustrate the geometric distinctions and trainability without extending the mathematical guarantees.

For trained layers, shared kernel parameters give an exact squared Gram norm and a Rademacher bound for a fixed RKHS ball. Fixing a trained prefix yields an exact pullback RKHS for the next layer, with a quotient norm. For positive bias, universality holds precisely when the continuous prefix is injective. This gives a local geometric description of composition, but not a uniform capacity bound over learned prefixes. Obtaining useful parameter-independent capacity control across depth remains open.

AI Use Statement

In this work, generative AI tools were used to assist with manuscript formatting and conversion from the NeurIPS style to the ICLR 2027 style, to implement methods, and to identify relevant literature for the scientific content of this paper. We did not use generative AI tools to generate mathematical claims, formulate proofs, design experiments, generate synthetic data, or interpret results. The authors reviewed all AI-assisted work and remain responsible for all claims, proofs, data, code, and artifacts.

Ethics Statement

This paper develops mathematical foundations for a kernel construction and contains no human-subject research, personal data, sensitive datasets, or high-risk model release. We are not aware of an ethics concern requiring a separate mitigation beyond the standard responsible-release and citation practices described in the manuscript.

Reproducibility Statement

Assumptions and proofs appear in Appendices A–E; diagnostic protocols, scripts, and records are described in Appendix F. Appendix G maps the artifacts and distinguishes reconstruction of the reported LM aggregates from complete historical reruns: checkpoint–evaluation linkage, code and environment revisions, and the evaluation-pass selection rationale remain incomplete.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Bastian Bohn, Michael Griebel, and Christian Rieger. A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, 20(64):1–32, 2019.

Nicolas Boullé, Yuji Nakatsukasa, and Alex Townsend. Rational neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2(3):321–355, 1988.

Martin D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, 2003.

Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, 2009.

Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational kernels: Theory and algorithms. *Journal of Machine Learning Research*, 5:1035–1062, 2004.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, 2009.

Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory (COLT)*, pp. 907–940, 2016.

Marc G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. In *Advances in Neural Information Processing Systems*, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Jordan Hoffmann et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in Neural Information Processing Systems*, 2014.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209: 415–446, 1909.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Alejandro Molina, Patrick Schramowski, and Kristian Kersting. Padé activation units: End-to-end learning of flexible activation functions in deep networks. In *International Conference on Learning Representations (ICLR)*, 2020.

John Moody and Christian J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Christopher J. Paciorek and Mark J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006.

Jooyoung Park and Irwin W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2016.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations (ICLR)*, 2020.

Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Alex J. Smola, Zoltán L. Óvári, and Robert C. Williamson. Regularization with dot-product kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Matus Telgarsky. Neural networks and rational functions. In *International Conference on Machine Learning (ICML)*, 2017.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

David Vernon Widder. *The Laplace Transform*. Princeton University Press, 1941.

Andrew Gordon Wilson and Ryan P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning (ICML)*, 2013.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In *AISTATS*, 2016.

Appendix guide. Appendix A collects notation, RKHS background, and related work. The mathematical development then follows three questions: what distance retains (Appendix B), what alignment changes (Appendix C), and how the atom families represent and approximate functions (Appendix D). Appendix E collects layer norms, capacity, and composition. Experimental protocols are collected separately in Appendix F. For the three main theorems, the covariance proofs are in Appendix C.2, exact recovery and its sharpness proof are in Appendix D.1, and the quadratic approximation and coefficient-budget proofs are in Appendix D.3.

A Background, Notation, and Related Work

A.1 Preliminaries and Notation

This appendix collects the notation, standard kernel facts, and related work used in the paper. The notation below distinguishes a fixed-kernel RKHS from a finite atom span whose parameters may vary. Appendix A.2 reviews RKHS background, including kernel order, universality, and PSD closure properties; Appendix A.2.1 states the two external theorems invoked by label in the proofs. Appendix A.3 places the results in the kernel and neural-network literature.

Table 1: Quick reference for the notation. The table is grouped by category; longer definitions and standing assumptions follow.

Symbol	Definition
Domains and dimensions	
d, d_0, d_ℓ	ambient dimension; network input and layer-ℓ output dimensions
$\mathcal{X} \subset \mathbb{R}^d$	compact input domain
$\mathbb{S}^{d-1}$	unit sphere $\{\mathbf{u} : \|\mathbf{u}\| = 1\}$
B_R	closed ball $\{\mathbf{x} : \|\mathbf{x}\| \le R\}$
Kernels and atoms	
$D_\varepsilon(\mathbf{x}, \mathbf{w})$	$\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$ (regularized squared distance)
$g_\varepsilon(\mathbf{x}; \mathbf{w}, b)$	$(\mathbf{x}^\top \mathbf{w} + b)^2 / D_\varepsilon$ (biased Yat atom)
$k_{b,\varepsilon}, k_{\mathtt{E}}$	Yat kernel g_ε; unbiased $k_{0,\varepsilon}$
$h_\varepsilon, k^\varepsilon_{\mathrm{IMQ}}$	IMQ kernel $1/D_\varepsilon$
RKHS objects	
$\mathcal{H}_{b,\varepsilon}$, $\mathcal{H}_{\mathtt{E}}$, $\mathcal{H}^\varepsilon_{\mathrm{IMQ}}$	RKHSs of $k_{b,\varepsilon}$, $k_{\mathtt{E}}$, h_ε
$\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$	squared norm of $\sum_j \alpha_j k(\mathbf{w}_j, \cdot)$
$\preceq$	Loewner PSD order on kernels / Gram matrices
Spans of atoms	
$F^+_{\mathtt{E},\varepsilon}$	$\mathrm{span}\{g_\varepsilon(\cdot; \mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d,\, b > 0\}$
$F^+_{\mathtt{E}}, F^{\ge 0}_{\mathtt{E}}$	broader Yat spans over $b > 0$ resp. $b \ge 0$, $\varepsilon > 0$
$F_{\mathrm{IMQ},\varepsilon}$	$\mathrm{span}\{h_\varepsilon(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d\}$
$\mathcal{V}_{\mathrm{IMQ}}(\Lambda)$	IMQ class with $\sum \|a_j\| \le \Lambda$ and unrestricted centers
$\mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)$	coefficient-variation and center bounds: $\sum \|a_j\| \le \Lambda$, $\|\mathbf{v}_j\| \le W$
Operators	
$T_\infty F(\mathbf{u})$	$\lim_{r \to \infty} F(r\mathbf{u})$ (asymptotic-trace operator)
$\mathcal{Q}_2(\mathbb{S}^{d-1})$	quadratic forms $\mathbf{u} \mapsto \mathbf{u}^\top \mathbf{A} \mathbf{u}$ on the sphere
Network and pullback	
$\Phi_\ell, T_\ell, m_\ell$	prefix map, layer map, number of kernel atoms
$z_{\ell,c}, \boldsymbol{\alpha}_{\ell,c}$	layer-ℓ output coordinate c and its read-out vector
$\widetilde{k}_\ell(\mathbf{x}, \mathbf{x}')$	$k_\ell(\Phi_{\ell-1}(\mathbf{x}), \Phi_{\ell-1}(\mathbf{x}'))$ (pullback)

Standing assumptions. Unless stated otherwise, $d \geq 1$, $\mathcal{X} \subset \mathbb{R}^d$ is compact, $b \geq 0$, $\varepsilon > 0$. The sphere is $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ and the closed ball $B_R = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq R\}$.

Kernels and atoms. We use the following notation for the offset squared distance, Yat atoms and kernels, and the IMQ kernel:

$$\begin{aligned} D_\varepsilon(\mathbf{x}, \mathbf{w}) &:= \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon, \\ \text{biased Yat atom:} \quad g_\varepsilon(\mathbf{x}; \mathbf{w}, b) &:= (\mathbf{x}^\top \mathbf{w} + b)^2 / D_\varepsilon(\mathbf{x}, \mathbf{w}), \\ \text{Yat kernel:} \quad k_{b,\varepsilon}(\mathbf{w}, \mathbf{x}) &:= g_\varepsilon(\mathbf{x}; \mathbf{w}, b), \quad k_{\mathbb{E}}(\mathbf{w}, \mathbf{x}) := k_{0,\varepsilon}(\mathbf{w}, \mathbf{x}), \\ \text{IMQ kernel:} \quad h_\varepsilon(\mathbf{x}, \mathbf{w}) &:= k^\varepsilon_{\mathrm{IMQ}}(\mathbf{x}, \mathbf{w}) := D_\varepsilon(\mathbf{x}, \mathbf{w})^{-1}. \end{aligned}$$

RKHS objects. $\mathcal{H}_{b,\varepsilon}$, $\mathcal{H}_{\mathbb{E}}$, $\mathcal{H}_{h_\varepsilon} = \mathcal{H}^\varepsilon_{\mathrm{IMQ}}$ denote the RKHSs of $k_{b,\varepsilon}$, $k_{\mathbb{E}}$, h_ε respectively. The Loewner PSD order is denoted $\preceq$. The reproducing property and existence theorem are recalled in Appendix A.2.

The squared norm of a finite expansion $f = \sum_j \alpha_j k(\mathbf{w}_j, \cdot)$ in the global RKHS on $\mathbb{R}^d$ is $\|f\|^2_{\mathcal{H}_k} = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$, where $\mathbf{K}_{ij} := k(\mathbf{w}_i, \mathbf{w}_j)$. The identity also holds with repeated centers; the Gram matrix may then be singular. Restriction to $\mathcal{X}$ gives a quotient norm, which can be smaller when centers lie outside $\mathcal{X}$.

Spans of atoms. The finite spans used in the paper are

$$\begin{aligned} F^+_{\mathbb{E},\varepsilon} &:= \mathrm{span}\{g_\varepsilon(\cdot; \mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d,\ b > 0\}, \\ F^+_{\mathbb{E}} &:= \mathrm{span}\{k_{b,\varepsilon}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d,\ b > 0,\ \varepsilon > 0\}, \\ F^{\geq 0}_{\mathbb{E}} &:= \mathrm{span}\{k_{b,\varepsilon}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d,\ b \geq 0,\ \varepsilon > 0\}, \\ F_{\mathrm{IMQ},\varepsilon} &:= \mathrm{span}\{k^\varepsilon_{\mathrm{IMQ}}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d\}. \end{aligned}$$

The subscript ε indicates a fixed distance offset. In $F^+_{\mathbb{E},\varepsilon}$ the bias may vary across atoms; in $F^+_{\mathbb{E}}$ and $F^{\geq 0}_{\mathbb{E}}$ both bias and distance offset may vary. These spans are distinct from the RKHS of a single fixed kernel.

The coefficient-variation class and its bounded-center subclass are

$$\begin{aligned} \mathcal{V}_{\mathrm{IMQ}}(\Lambda) &:= \Big\{ \textstyle\sum_j a_j h_\varepsilon(\cdot, \mathbf{v}_j) : \sum_j |a_j| \leq \Lambda,\ \mathbf{v}_j \in \mathbb{R}^d \Big\}, \\ \mathcal{R}_{\mathrm{IMQ}}(\Lambda, W) &:= \Big\{ \textstyle\sum_j a_j h_\varepsilon(\cdot, \mathbf{v}_j) : \sum_j |a_j| \leq \Lambda,\ \|\mathbf{v}_j\| \leq W \Big\}, \end{aligned}$$

where the unrestricted class $\mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ is used in the bounded-domain coefficient-budget obstruction; the center restriction in $\mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)$ is used only for exterior-shell decay.

Operators on the spans. The directional asymptotic trace is defined by radial limits:

$$T_\infty F(\mathbf{u}) \ := \ \lim_{r \to \infty} F(r\mathbf{u}), \qquad \mathbf{u} \in \mathbb{S}^{d-1},$$

on the linear subspace $\mathcal{D}_{T_\infty} \subset C(\mathbb{R}^d)$ where the limit exists for every direction. The image of T_∞ on $F^{\geq 0}_{\mathbb{E}}$ is the space of quadratic forms restricted to $\mathbb{S}^{d-1}$, $\mathcal{Q}_2(\mathbb{S}^{d-1}) := \{\mathbf{u} \mapsto \mathbf{u}^\top \mathbf{A} \mathbf{u} : \mathbf{A} \in \mathrm{Sym}(d)\}$.

Network notation. For a depth-L Yat stack with input $\mathbf{x} \in X \subset \mathbb{R}^{d_0}$, output dimensions d_ℓ, and m_ℓ kernel atoms per layer,

$$\Phi_0(\mathbf{x}) = \mathbf{x}, \qquad \Phi_\ell(\mathbf{x}) = T_\ell(\Phi_{\ell-1}(\mathbf{x})), \qquad [T_\ell(\mathbf{z})]_c = \sum_{j=1}^{m_\ell} \alpha_{\ell,j,c}\, k_{b_\ell,\varepsilon_\ell}(\mathbf{w}_{\ell,j}, \mathbf{z}),$$

where $\Phi_{\ell-1}$ is the prefix map, T_ℓ the layer map, $1 \leq c \leq d_\ell$ a coordinate index, and $z_{\ell,c} := [\Phi_\ell]_c$ the layer-ℓ output coordinate. Write $k_\ell := k_{b_\ell,\varepsilon_\ell}$. The pulled-back kernel on X is $\widetilde{k}_\ell(\mathbf{x}, \mathbf{x}') := k_\ell(\Phi_{\ell-1}(\mathbf{x}), \Phi_{\ell-1}(\mathbf{x}'))$.

Conventions. $\mathbb{S}^{d-1}$ carries normalized uniform surface measure (total mass one) where unspecified. Uniform norms on sets, also denoted L^∞ in the approximation statements, mean pointwise suprema rather than measure-dependent essential suprema. $\mathrm{Sym}(d)$ is the space of real $d \times d$ symmetric matrices; $\mathrm{vec}(\cdot)$ is column-major matrix vectorisation; $\langle \cdot, \cdot \rangle$ without subscript is the Euclidean inner product on $\mathbb{R}^d$. The Tifinagh letter $\mathtt{E}$ (Unicode U+2D5F) is used as the kernel-specific glyph $k_{\mathtt{E}}$.

A.2 RKHS Background and External Results

Reproducing kernel Hilbert spaces. A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semidefinite (PSD)* if $\sum_{i,j} c_i c_j \, k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ for every finite set $\{\mathbf{x}_i\} \subset \mathcal{X}$ and coefficients $\{c_i\} \subset \mathbb{R}$. Every continuous PSD k on a compact metric space $\mathcal{X}$ is a *Mercer kernel*. More precisely, fix a finite Borel measure ρ with full support on $\mathcal{X}$. The associated integral operator on $L^2(\rho)$ is compact, self-adjoint, and positive; Mercer's theorem gives $k(\mathbf{x}, \mathbf{y}) = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y})$ with $\lambda_i \geq 0$, $\{\phi_i\}$ orthonormal in $L^2(\rho)$, and absolute uniform convergence on $\mathcal{X} \times \mathcal{X}$ (Mercer, 1909).

The Moore–Aronszajn theorem (Aronszajn, 1950) associates to every PSD k a unique *reproducing kernel Hilbert space (RKHS)* $\mathcal{H}_k$ whose reproducing property $f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k}$ holds for all $f \in \mathcal{H}_k$.

Kernel order and sums. Two RKHS properties are used throughout. First, if $c > 0$ and $c^2 k_1 \preceq k_2$ in the Loewner PSD order (meaning $k_2 - c^2 k_1$ is PSD), then $\mathcal{H}_{k_1} \subseteq \mathcal{H}_{k_2}$ with $\|f\|_{\mathcal{H}_{k_2}} \leq c^{-1} \|f\|_{\mathcal{H}_{k_1}}$ (Paulsen & Raghupathi, 2016). Second, the RKHS sum rule: if $k = k_1 + k_2$ then $\mathcal{H}_k = \mathcal{H}_{k_1} + \mathcal{H}_{k_2}$ with squared infimal-convolution norm $\|f\|^2_{\mathcal{H}_k} = \inf_{f_1 + f_2 = f} (\|f_1\|^2_{\mathcal{H}_{k_1}} + \|f_2\|^2_{\mathcal{H}_{k_2}})$.

Universality, characteristicness, and strict positive definiteness. A continuous PSD kernel on compact $\mathcal{X}$ is *universal* if its RKHS is dense in $C(\mathcal{X})$ (Micchelli et al., 2006; Steinwart & Christmann, 2008). We say k is *characteristic* if the kernel mean embedding $\mu \mapsto \int k(\cdot, \mathbf{x}) \, d\mu(\mathbf{x})$ is injective on finite signed Borel measures on $\mathcal{X}$ (this finite-signed formulation is the one we use throughout). A kernel is *strictly positive definite (SPD)* if for every $n \geq 1$ and every set of pairwise distinct points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the Gram matrix is positive definite. The implications we use for continuous kernels on compact $\mathcal{X}$ are

$$\text{universal} \implies \text{characteristic} \implies \text{SPD},$$

where the first arrow is Sriperumbudur et al. (2011) and the second follows because if $\mathbf{c}^\top \mathbf{K} \mathbf{c} = 0$ for distinct nodes then the atomic measure $\mu = \sum_i c_i \delta_{\mathbf{x}_i}$ has zero kernel mean embedding.

The measure convention matters: under the finite-signed-measure definition above, characteristicness on a compact domain is equivalent to universality by a Hahn–Banach/Riesz argument. Injectivity restricted to probability measures is the weaker convention. We use the implication from universality to injectivity on finite signed measures. For a continuous universal kernel on a compact metric space, MMD_k metrizes weak convergence of probability measures (Sriperumbudur et al., 2010, Thm. 23). Applying this result directly through universality gives the additional consequence in Corollary 3.

Schur products and Laplace integrals. Two closure properties are used repeatedly. **(S)** The *Schur product theorem* (Horn & Johnson, 2012, Sec. 7.5): the entrywise (Hadamard) product of two PSD kernels is PSD.

(L) The Laplace representation $a^{-1} = \int_0^\infty e^{-ta} \, dt$ for $a > 0$ (Widder, 1941, Ch. IV): the IMQ kernel $h_\varepsilon(\mathbf{x}, \mathbf{w}) = (\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon)^{-1} = \int_0^\infty e^{-t\varepsilon} e^{-t\|\mathbf{x}-\mathbf{w}\|^2} \, dt$ is a nonnegative mixture of Gaussian PSD kernels, hence PSD.

Combining (S) and (L), any Schur product $p(\mathbf{x}, \mathbf{w}) \cdot h_\varepsilon(\mathbf{x}, \mathbf{w})$ with a PSD polynomial kernel p is PSD — this is the core of the Yat PSD proof. For the Yat numerator $p_b(\mathbf{x}, \mathbf{w}) = (\mathbf{x}^\top \mathbf{w} + b)^2$, the feature map $\Phi_b(\mathbf{x}) = (\mathbf{x} \otimes \mathbf{x}, \sqrt{2b}\,\mathbf{x}, b)$ requires $b \geq 0$ for $\sqrt{2b}$ to be real; the sign condition is necessary, as Theorem 6 exhibits a 2×2 counterexample for every $b < 0$.

A.2.1 External theorems used in the proofs

The following two results supply product-kernel containment and IMQ universality in the proofs of Sections 2–5. The preceding background gives the kernel-order and PSD closure rules used alongside them.

Theorem 4 (Product-kernel RKHS containment (Steinwart & Christmann, 2008, Lemma 4.6 and Thm. 4.21))**.** *Let k_1, k_2 be continuous PSD kernels on $\mathcal{X}$, and let $\mathcal{H}_1, \mathcal{H}_2$ be their RKHSs. Then the Schur product $k = k_1 \cdot k_2$ is a continuous PSD kernel with RKHS $\mathcal{H}_k$ satisfying: for every $f_1 \in \mathcal{H}_1$ and $f_2 \in \mathcal{H}_2$, the pointwise product $f_1 \cdot f_2 \in \mathcal{H}_k$, with $\|f_1 \cdot f_2\|_{\mathcal{H}_k} \le \|f_1\|_{\mathcal{H}_1} \|f_2\|_{\mathcal{H}_2}$.*

The product feature map is $\Phi(x) = k_1(\cdot, x) \otimes k_2(\cdot, x)$ in $\mathcal{H}_1 \widehat{\otimes} \mathcal{H}_2$ (Lemma 4.6). Theorem 4.21 identifies its RKHS norm with the minimum representing-vector norm. The vector $f_1 \otimes f_2$ represents $f_1 f_2$ and has norm $\|f_1\|_{\mathcal{H}_1} \|f_2\|_{\mathcal{H}_2}$, giving the stated bound.

Theorem 5 (Micchelli IMQ universality (Micchelli et al., 2006, Thm. 17))**.** *For every $\varepsilon > 0$ and every compact $\mathcal{X} \subset \mathbb{R}^d$, the IMQ span $F_{\mathrm{IMQ},\varepsilon} = \mathrm{span}\{(\|\cdot - \mathbf{w}\|^2 + \varepsilon)^{-1} : \mathbf{w} \in \mathbb{R}^d\}$ is dense in $C(\mathcal{X})$ in the uniform norm; equivalently, the IMQ kernel is universal on every compact $\mathcal{X} \subset \mathbb{R}^d$.*

A.3 Related Work

Table 2: Geometry and guarantees of kernel-valued hidden units. "Fixed UAT" means universality of one fixed kernel on every compact domain, with the zero-bias exception stated in its row. "Distance factor" denotes a radial factor based on $\|\mathbf{x} - \mathbf{w}\|$, not decay of the complete section. "Alignment" denotes an explicit inner-product factor in addition to any distance factor. "Nonzero trace" means that a section with a nonzero center can have a finite limit in every direction, with the resulting function of direction not identically zero. The exact squared Gram norm is shared by every fixed kernel expansion and is therefore not used as a distinguishing column.

Primitive	Distance factor	Alignment	Fixed UAT	Each fixed section bounded	Nonzero trace	IMQ relation
Yat ($b > 0$)	✓	✓	✓	✓	✓	contains IMQ RKHS
Yat ($b = 0$)	✓	✓	iff $0 \notin \mathcal{X}$[‡]	✓	✓	—
Variable-b Yat span	✓	✓	family UAT	✓	✓	three-atom exact[†]
IMQ	✓	—	✓	✓	—	baseline
Gaussian RBF	✓	—	✓	✓	—	—
Quadratic poly.	—	✓	—	—	—	—
Univ. dot-product[§]	—	✓	✓	—	—	—

[†]The variable-bias row fixes $\varepsilon > 0$. Exact recovery uses three atoms with different positive biases; fixed $b > 0$ gives IMQ-RKHS containment. [‡]Proposition 4. [§]Example: $e^{\gamma \mathbf{x}^\top \mathbf{w}}$, $\gamma > 0$ (Smola et al., 2000). Along $\mathbf{x} = r\mathbf{u}$, this section tends to infinity, zero, or one according as $\mathbf{u}^\top \mathbf{w}$ is positive, negative, or zero; it has no finite trace in every direction when $\mathbf{w} \neq 0$. Quadratic polynomial sections likewise diverge in directions with nonzero alignment. All listed fixed kernels are PSD in their stated parameter regimes. The variable-bias row describes a family of atoms, rather than one Mercer kernel.

RBF and learned-center kernel networks. Radial-basis-function networks (Broomhead & Lowe, 1988; Moody & Darken, 1989) train finite expansions of locally-tuned radial kernels and are universal approximators (Park & Sandberg, 1991). Related classical MLP approximation theory is surveyed by Pinkus (1999); native-space theory and IMQ universality are developed in Wendland (2004).

With a shared bandwidth, an IMQ or Gaussian RBF layer is already a finite learned-center RKHS expansion admitting the same $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$ norm identity and diagonal-based Rademacher

argument as Yat. The alignment factor $(\mathbf{w}^\top \mathbf{x} + b)^2$ changes the behavior of the sections: a Yat atom with a nonzero center has a nonzero directional far-field trace, whereas finite IMQ and Gaussian RBF expansions decay along every ray. In addition, $b > 0$ supplies a constant-coordinate feature in the polynomial factor. Kernel-order domination then establishes universality for each fixed positive bias and distance offset parameter; varying the bandwidth is not needed.

Neural kernels and arc-cosine constructions. Cho & Saul (2009) introduced arc-cosine kernels associated with threshold activations. Both arc-cosine and Yat constructions relate neural units to positive-semidefinite kernels. Here the trainable center is itself one argument of the symmetric kernel, and each hidden unit is a finite rational section. The comparisons developed below concern this explicit section's kernel order, RKHS sum decomposition, and finite-difference identity, rather than an identification with an arc-cosine kernel.

Kernel machines and product kernels. Polynomially modulated radial kernels and product or nonstationary kernel constructions are classical in kernel machines and Gaussian processes (Schölkopf & Smola, 2002; Steinwart & Christmann, 2008). The tensor-product construction and feature-space characterization of Steinwart–Christmann (Steinwart & Christmann, 2008, Lemma 4.6 and Thm. 4.21) provide one route to fixed-kernel universality; the kernel-order argument provides another. Our contribution is not product-kernel validity itself. We analyze the resulting rational kernel through sharp IMQ domination, directional traces and strict global containment, the $C \circ C$ covariance lift, optimal three-atom exact recovery, and coefficient-budget separation. The layer-local Gram-norm identity is a standard RKHS consequence used to interpret trained parameters.

Deep composition and kernel viewpoints. For standard scalar-activation MLPs, layer-level capacity is controlled through spectral or Lipschitz surrogates on the weight matrix (Bartlett et al., 2017; Neyshabur et al., 2018), which are function-level bounds rather than Hilbert-space objects attached to individual units.

The conjugate-kernel construction of Daniely et al. (2016) relates neural networks and random initial representations to compositional kernel spaces. Deep kernel learning (Wilson et al., 2016) combines a learned feature extractor with a Gaussian-process kernel. Both connect feature composition to kernels, but Theorem 13 conditions on a fixed prefix and characterizes the resulting RKHS norm; it is not an initialization or optimization guarantee.

The representer theorem of Bohn et al. (2019) gives finite-dimensional representations for optimization over concatenated RKHS functions. Our pullback result instead identifies the space and quotient norm after a prefix is fixed. Convolutional kernel networks (Mairal et al., 2014) learn finite approximations to kernel feature maps. Here a shared-parameter Yat layer is already an exact finite sum of sections of the specified kernel; no approximation to that kernel's feature map is required for its Gram-norm identity.

RKHS penalties have also been used to regularize ordinary deep networks by matching their functions to kernel-induced smoothness (Bietti et al., 2019). That line attaches an RKHS regularizer to a network function. Here the architectural primitive itself is a kernel section, so a shared-parameter layer has an exact finite-expansion norm and the Yat-specific finite-difference and directional identities remain accessible after training. For this kernel, the explicit kernel inequality, minimal three-atom IMQ representation, and directional asymptotic trace complement the general compositional framework.

RKHS generalization bounds. The single-layer Rademacher bound (Theorem 12) follows the RKHS-ball Rademacher framework of Bartlett & Mendelson (2002). The relationships among universality, injective kernel mean embeddings, and strict positive definiteness depend on the definition of characteristicness; throughout we use injectivity on finite signed measures (Sriperumbudur et al., 2011). For Yat, the input-dependent diagonal $k_{b,\varepsilon}(\mathbf{x}, \mathbf{x}) = (\|\mathbf{x}\|^2 + b)^2/\varepsilon$ makes b and ε explicit in the bound, unlike the constant diagonal of the Gaussian (1) or IMQ (ε^{-1}) kernels.

Approximation rates and shallow-network separations. The coefficient-budget obstruction in Theorem 3 is distinct from classical neural-network depth separations. Eldan & Shamir (2016) constructs an approximately radial target that is efficiently represented by a depth-three network but requires exponential width at depth two. Our comparison instead concerns IMQ coefficient budgets for approximating positive-semidefinite quadratic targets. Bach (2017) studies approximation in convex neural-network function classes controlled by variation-type norms. Those results provide methodological context, but neither is invoked to prove our IMQ budget inequality. The Barron-space MLP capacity literature (E et al., 2019) provides a related functional class for two-layer networks. Yat instead places each shared-parameter layer in a fixed RKHS.

Rational and ratio-form kernels. The term "rational kernel" is also used for the sequence kernels of Cortes et al. (2004), which are constructed by finite-state operations; that framework is distinct from the rational function on $\mathbb{R}^d$ considered here. Rational scalar activations are analyzed by Telgarsky (2017), Molina et al. (2020), and Boullé et al. (2020). Those models remain scalar functions of $\mathbf{w}^\top \mathbf{x}$ rather than sections of a fixed kernel on the joint input–parameter domain. Normalized polynomial kernels (Burges, 1998; Schölkopf & Smola, 2002) divide a polynomial kernel by its diagonal normalization, whereas Yat uses the regularized distance $\|\mathbf{x}-\mathbf{w}\|^2+\varepsilon$. The native-space theory of IMQ kernels is developed in Wendland (2004); Buhmann (2003).

Non-stationary and anisotropic kernels. The Yat kernel is non-stationary in the sense that $k_{b,\varepsilon}(\mathbf{w},\mathbf{x})$ depends on $(\mathbf{w},\mathbf{x})$ rather than only on $\mathbf{x}-\mathbf{w}$. It is jointly orthogonally invariant: $k_{b,\varepsilon}(Q\mathbf{w},Q\mathbf{x}) = k_{b,\varepsilon}(\mathbf{w},\mathbf{x})$ for every orthogonal Q. A fixed-center section is directionally selective through $(\mathbf{x}^\top\mathbf{w}+b)^2$, with preferred direction $\mathbf{w}$. This places Yat in a literature on non-stationary covariance constructions in Gaussian processes and kernel machines: Genton (2001) surveys the design space of kernel families including non-stationary and anisotropic constructions, Paciorek & Schervish (2006) develop a class of non-stationary covariance functions parameterised by spatially varying length scales. In contrast, the spectral mixture kernels of Wilson & Adams (2013) model a stationary covariance through a Gaussian mixture spectral density; their frequency structure is distinct from Yat's non-stationary polynomial alignment factor. The Yat kernel uses a single rational expression rather than a spectral mixture or a spatially varying length-scale field. Its algebra yields the exact IMQ embedding in Theorem 2. The closed-form layer-local RKHS norm is a separate, generic property of shared-bandwidth Mercer-section layers.

B What Distance Retains: Proofs and Kernel Bounds

This appendix supplies the fixed-kernel guarantees used in Section 2. We first establish positive semidefiniteness and the sharp IMQ domination constant. We then prove universality, including its exact zero-bias boundary, and derive the measure-embedding and Gram-matrix consequences. The final subsection distinguishes bounded-domain estimates, global bounds for individual sections, and the diagonal used in the capacity bound.

Proposition 2 (Kernel validity and inherited approximation guarantees)**.** *Fix $b \geq 0$ and $\varepsilon > 0$.* (a) Validity and order. *The kernel $k_{b,\varepsilon}$ is symmetric positive semidefinite on $\mathbb{R}^d$ and continuous Mercer on every compact restriction. It satisfies $b^2 h_\varepsilon \preceq k_{b,\varepsilon}$.*

(b) RKHS embedding. *For $b > 0$,*

$$\mathcal{H}_{h_\varepsilon} \subseteq \mathcal{H}_{b,\varepsilon}, \qquad \|f\|_{\mathcal{H}_{b,\varepsilon}} \leq b^{-1}\|f\|_{\mathcal{H}_{h_\varepsilon}}.$$

(c) Compact-domain guarantees. *For $b > 0$, on every compact $\mathcal{X} \subset \mathbb{R}^d$ the fixed kernel is universal, its mean embedding is injective on finite signed Borel measures, and its MMD metrizes weak convergence of Borel probability measures. Its Gram matrix on any finite set of distinct points is strictly positive definite.*

(d) Zero-bias boundary. *At $b = 0$, the kernel is universal on a nonempty compact $\mathcal{X}$ if and only if $\mathbf{0} \notin \mathcal{X}$.*

B.1 Positive semidefiniteness and Loewner domination

Theorem 6 (PSD and Mercer for $b \geq 0$)**.** *For every $b \geq 0$, $\varepsilon > 0$, the kernel $k_{b,\varepsilon}$ in (1) is symmetric and positive semidefinite on $\mathbb{R}^d \times \mathbb{R}^d$, and on every compact $\mathcal{X} \subset \mathbb{R}^d$ the restriction is continuous and Mercer. By Moore–Aronszajn there exists a unique RKHS $\mathcal{H}_{b,\varepsilon}$ with reproducing kernel $k_{b,\varepsilon}$. The sign condition is essential on the full domain: for every $b < 0$ and every $d \geq 1$, $k_{b,\varepsilon}$ is not PSD on $\mathbb{R}^d$.*

Proof. For $b \geq 0$, we prove that both factors are PSD and apply the Schur product theorem. We then give counterexamples for negative bias. Factor $k_{b,\varepsilon} = p \cdot h_\varepsilon$, where $p(\mathbf{x}, \mathbf{w}) := (\mathbf{x}^\top \mathbf{w} + b)^2$ and $h_\varepsilon(\mathbf{x}, \mathbf{w}) := (\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon)^{-1}$.

Polynomial factor. Append the same coordinate $\sqrt{b}$ to both inputs: define $\tilde{\mathbf{x}} := (\mathbf{x}, \sqrt{b}) \in \mathbb{R}^{d+1}$ and $\tilde{\mathbf{w}} := (\mathbf{w}, \sqrt{b}) \in \mathbb{R}^{d+1}$ (well-defined because $b \geq 0$). Then $\tilde{\mathbf{x}}^\top \tilde{\mathbf{w}} = \mathbf{x}^\top \mathbf{w} + b$, so $p(\mathbf{x}, \mathbf{w}) = (\tilde{\mathbf{x}}^\top \tilde{\mathbf{w}})^2$ is the square of a linear kernel in extended feature space. The linear kernel is PSD; by the Schur product theorem its Hadamard square is also PSD, so p is PSD.

An explicit feature map gives the same conclusion: $\Phi_b(\mathbf{x}) = (\mathrm{vec}(\mathbf{x}\mathbf{x}^\top), \sqrt{2b}\,\mathbf{x}, b) \in \mathbb{R}^{d^2+d+1}$ satisfies $\langle \Phi_b(\mathbf{x}), \Phi_b(\mathbf{w}) \rangle = (\mathbf{x}^\top \mathbf{w})^2 + 2b\,\mathbf{x}^\top \mathbf{w} + b^2 = (\mathbf{x}^\top \mathbf{w} + b)^2 = p(\mathbf{x}, \mathbf{w})$. This feature map is real-valued for $b \geq 0$. The counterexamples below establish that the sign condition is necessary for the full kernel, not merely for this feature representation.

IMQ factor. By the Laplace identity $a^{-1} = \int_0^\infty e^{-ta}\, dt$ for $a > 0$,

$$h_\varepsilon(\mathbf{x}, \mathbf{w}) \;=\; \int_0^\infty e^{-t\varepsilon}\, e^{-t\|\mathbf{x}-\mathbf{w}\|^2}\, dt. \tag{9}$$

For each $t > 0$ the Gaussian $e^{-t\|\mathbf{x}-\mathbf{w}\|^2}$ is a PSD kernel and $e^{-t\varepsilon} \geq 0$. A nonnegative mixture of PSD kernels is PSD, so h_ε is PSD.

Schur product. By the Schur product theorem (Horn & Johnson, 2012), the entry-wise product of two PSD matrices is PSD. Applied to the Gram matrices of p and h_ε on an arbitrary finite point set, this shows $k_{b,\varepsilon} = p \cdot h_\varepsilon$ is PSD on $\mathbb{R}^d$.

Continuity on compact $\mathcal{X}$. The denominator $\|\mathbf{x}-\mathbf{w}\|^2 + \varepsilon$ is bounded below by $\varepsilon > 0$, so $k_{b,\varepsilon}$ is a ratio of a continuous numerator and a strictly positive continuous denominator. It is therefore a continuous Mercer kernel on $\mathcal{X} \times \mathcal{X}$. Moore–Aronszajn gives the unique RKHS.

Negative bias: a concrete example. Take $b = -1$, $d = 2$, nodes $\mathbf{x}_1 = \mathbf{e}_1$, $\mathbf{x}_2 = \mathbf{e}_2$. The numerator Gram is

$$[(\mathbf{x}_i^\top \mathbf{x}_j + b)^2] \;=\; \begin{pmatrix} (1-1)^2 & (0-1)^2 \\ (0-1)^2 & (1-1)^2 \end{pmatrix} \;=\; \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

with eigenvalues $\{1, -1\}$, so the numerator is indefinite. The denominator at these nodes is $D_\varepsilon(\mathbf{x}_i, \mathbf{x}_j) \in \{\varepsilon, 2 + \varepsilon\}$, both positive. The full kernel Gram matrix therefore has zero diagonal and off-diagonal entries $(2+\varepsilon)^{-1}$, giving the negative eigenvalue $-(2+\varepsilon)^{-1}$; hence $k_{-1,\varepsilon}$ is not PSD.

Negative bias: all dimensions and magnitudes. For the full necessity statement, take any $b < 0$, $d \geq 1$, $\mathbf{x} = \sqrt{-b}\,\mathbf{e}_1$, and $\mathbf{y} = \mathbf{0}$. Then

$$k_{b,\varepsilon}(\mathbf{x}, \mathbf{x}) = 0, \qquad k_{b,\varepsilon}(\mathbf{x}, \mathbf{y}) = \frac{b^2}{-b + \varepsilon} > 0.$$

The two-point Gram determinant is $-[b^2/(-b + \varepsilon)]^2 < 0$. Thus no negative bias gives a PSD kernel on all of $\mathbb{R}^d$, regardless of $\varepsilon > 0$. □

The PSD factorization also identifies an IMQ component inside the kernel. Keeping that component and grouping the remaining PSD terms gives the following norm-controlled inclusion.

Theorem 7 (Kernel-order domination). *For every $b \geq 0$, $\varepsilon > 0$,*

$$b^2\, h_\varepsilon \preceq k_{b,\varepsilon}$$

in the Loewner PSD order on kernels. For $b > 0$, Aronszajn's inclusion theorem (Paulsen & Raghupathi, 2016) gives

$$\mathcal{H}_{h_\varepsilon} \subseteq \mathcal{H}_{b,\varepsilon}, \qquad \|f\|_{\mathcal{H}_{b,\varepsilon}} \leq \frac{1}{b}\|f\|_{\mathcal{H}_{h_\varepsilon}} \quad (b > 0).$$

Equivalently, for $b > 0$, $B_{\mathcal{H}_{h_\varepsilon}}(R) \subseteq B_{\mathcal{H}_{b,\varepsilon}}(R/b)$. At $b = 0$ only the kernel inequality is asserted, not IMQ-RKHS containment.

Proof of Theorem 7. By Theorem 6, $k_{b,\varepsilon}$ defines the RKHS $\mathcal{H}_{b,\varepsilon}$. Compute

$$k_{b,\varepsilon} - b^2 h_\varepsilon = 2b(\mathbf{x}^\top \mathbf{w})h_\varepsilon + (\mathbf{x}^\top \mathbf{w})^2 h_\varepsilon.$$

The linear and quadratic dot-product kernels are PSD. Since $2b \geq 0$ and h_ε is PSD, both terms on the right are PSD Schur products; their sum is PSD by the Schur product theorem (Horn & Johnson, 2012, Sec. 7.5). Thus $b^2 h_\varepsilon \preceq k_{b,\varepsilon}$, and Aronszajn's inclusion theorem yields the stated RKHS containment and norm bound. At $b = 0$ the kernel inequality becomes $0 \preceq k_{0,\varepsilon}$; the nontrivial IMQ-RKHS inclusion begins at $b > 0$. □

Sharp constant and zero-bias endpoint. The constant b^2 is optimal on $\mathbb{R}^d$. At $\mathbf{x} = \mathbf{0}$, $k_{b,\varepsilon}(\mathbf{0},\mathbf{0})/h_\varepsilon(\mathbf{0},\mathbf{0}) = b^2$; hence the 1×1 Gram matrix of $k_{b,\varepsilon} - c h_\varepsilon$ is negative for every $c > b^2$. Therefore

$$\sup\{c \geq 0 : c\, h_\varepsilon \preceq k_{b,\varepsilon}\} = b^2.$$

On a compact domain separated from the origin, the optimal domain-restricted constant may be larger.

The kernel and norm comparisons have different scalings as $b \to 0^+$. The domination constant b^2 compares kernels and their Gram matrices; the embedding constant $1/b$ in Theorem 7 compares RKHS norms. Thus estimates transferred through this embedding acquire a factor $1/b$ for norms, or $1/b^2$ for squared norms, and deteriorate as $b \downarrow 0$.

The endpoint $b = 0$ changes the RKHS. By Theorem 8, $\mathcal{H}_{0,\varepsilon} = \mathcal{H}_{k_2}$, while the radial and linear-alignment channels vanish. Moreover, every $f \in \mathcal{H}_{0,\varepsilon}$ satisfies $f(\mathbf{0}) = 0$, whereas the IMQ section $h_\varepsilon(\cdot, \mathbf{0})$ is nonzero there. Thus $\mathcal{H}_{h_\varepsilon} \not\subset \mathcal{H}_{0,\varepsilon}$, even though $\mathcal{H}_{h_\varepsilon} \hookrightarrow \mathcal{H}_{b,\varepsilon}$ for every $b > 0$. Positive bias gives universality on every compact domain; at zero bias the exact domain dependence is given below.

B.2 Fixed-kernel universality, characteristicness, and strict positive definiteness

For positive bias, kernel-order domination embeds the universal IMQ RKHS and therefore already gives universality. The alternative product-space proof below makes the role of the constant feature explicit and recovers the same norm bound. At zero bias, a nonvanishing quadratic feature replaces the constant feature precisely on compact domains that avoid the origin.

Proposition 3 (Fixed-kernel universality for $b_0 > 0$). *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and let $b_0 > 0$, $\varepsilon_0 > 0$. Then the RKHS $\mathcal{H}_{b_0,\varepsilon_0}$ of the fixed biased kernel $k_{b_0,\varepsilon_0}(\mathbf{w},\mathbf{x}) = (\mathbf{w}^\top \mathbf{x} + b_0)^2/(\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon_0)$ is dense in $C(\mathcal{X})$ in the uniform norm. Hence the single fixed kernel k_{b_0,ε_0} is universal.*

Proof. Write $k = p \cdot h$ with $h(\mathbf{x},\mathbf{w}) = (\|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon_0)^{-1}$ the IMQ kernel and $p(\mathbf{x},\mathbf{w}) = (\mathbf{x}^\top \mathbf{w} + b_0)^2$.

Step 1: the IMQ factor is universal. The RKHS $\mathcal{H}_h$ of h is dense in $C(\mathcal{X})$ for every compact $\mathcal{X}$ by the Micchelli universality theorem (Micchelli et al., 2006).

Step 2: the polynomial factor has a constant feature. The numerator kernel p has the explicit feature map $\Phi_{b_0}(\mathbf{x}) = (\mathrm{vec}(\mathbf{x}\mathbf{x}^\top), \sqrt{2b_0}\,\mathbf{x}, b_0) \in \mathbb{R}^{d^2+d+1}$, so $p(\mathbf{x}, \mathbf{w}) = \langle \Phi_{b_0}(\mathbf{x}), \Phi_{b_0}(\mathbf{w}) \rangle$. One coordinate of Φ_{b_0} is the nonzero constant b_0.

Step 3: the polynomial RKHS contains constants. The RKHS of p is $\mathcal{H}_p = \{\langle \mathbf{u}, \Phi_{b_0}(\cdot) \rangle : \mathbf{u} \in \mathbb{R}^{d^2+d+1}\}$. The dual coefficient $\mathbf{u}_1 = (0, \ldots, 0, 1/b_0)$ yields $\langle \mathbf{u}_1, \Phi_{b_0}(\mathbf{x}) \rangle = b_0 \cdot (1/b_0) = 1$ for every $\mathbf{x}$, so the constant function $\mathbf{1}_\mathcal{X} \in \mathcal{H}_p$ with $\|\mathbf{1}_\mathcal{X}\|_{\mathcal{H}_p} \le \|\mathbf{u}_1\| = 1/b_0$ (an upper bound because Φ_{b_0} is non-minimal, so the RKHS norm is the infimum over representers).

Step 4: $\mathcal{H}_k$ contains $\mathcal{H}_h$. By the product-kernel RKHS containment theorem (Steinwart & Christmann, 2008, Lemma 4.6 and Thm. 4.21) (also Theorem 4 above), the RKHS of $k = p \cdot h$ contains every pointwise product $f_1 \cdot f_2$ with $f_1 \in \mathcal{H}_h$ and $f_2 \in \mathcal{H}_p$, and $\|f_1 \cdot f_2\|_{\mathcal{H}_k} \le \|f_1\|_{\mathcal{H}_h} \|f_2\|_{\mathcal{H}_p}$. Take $f_2 = \mathbf{1}_\mathcal{X} \in \mathcal{H}_p$ from Step 3, with $\|\mathbf{1}_\mathcal{X}\|_{\mathcal{H}_p} \le 1/b_0$. Then for every $f_1 \in \mathcal{H}_h$ the pointwise product $f_1 \cdot \mathbf{1}_\mathcal{X} = f_1$ lies in $\mathcal{H}_k$, and $\|f_1\|_{\mathcal{H}_k} \le \|f_1\|_{\mathcal{H}_h} \cdot (1/b_0)$. Thus the embedding $\mathcal{H}_h \hookrightarrow \mathcal{H}_k$ has operator norm $\le 1/b_0$, recovering the Loewner bound $\|f\|_{\mathcal{H}_{b_0,\varepsilon_0}} \le b_0^{-1} \|f\|_{\mathcal{H}_h}$ of Theorem 7. The kernel-order proof gives the same bound directly.

Step 5: density. Let $g \in C(\mathcal{X})$ and $\eta > 0$. By Step 1, there exists $f \in \mathcal{H}_h$ with $\|g - f\|_\infty < \eta$. Step 4 gives $f \in \mathcal{H}_k$. Hence $\mathcal{H}_k$ is dense in $C(\mathcal{X})$. □

At zero bias, the constant feature is unavailable. The next argument uses $\|\mathbf{x}\|^2$ instead; whether this function vanishes on the domain determines the universality boundary.

Proposition 4 (Exact zero-bias compact-domain boundary)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty compact set and fix $\varepsilon > 0$. The RKHS $\mathcal{H}_{0,\varepsilon}|_\mathcal{X}$ is dense in $C(\mathcal{X})$ if and only if $\mathbf{0} \notin \mathcal{X}$.*

Proof of Proposition 4. Domains containing the origin. If $\mathbf{0} \in \mathcal{X}$, every section $k_{0,\varepsilon}(\mathbf{w}, \cdot)$ vanishes at $\mathbf{0}$, and continuity of point evaluation implies that every member of the RKHS does too; the constant function cannot be approximated uniformly.

Domains avoiding the origin. Suppose $\mathbf{0} \notin \mathcal{X}$ and put $q(\mathbf{x}) = \|\mathbf{x}\|^2$. The homogeneous quadratic kernel $(\mathbf{x}^\top \mathbf{w})^2$ has q in its finite-dimensional RKHS. Since $k_{0,\varepsilon} = ((\mathbf{x}^\top \mathbf{w})^2) h_\varepsilon$, the tensor-product feature map shows that qh belongs to $\mathcal{H}_{0,\varepsilon}$ whenever $h \in \mathcal{H}_{h_\varepsilon}$. For any $f \in C(\mathcal{X})$, the function f/q is continuous because q is bounded below on $\mathcal{X}$. Universality of the IMQ kernel supplies $h_n \in \mathcal{H}_{h_\varepsilon}$ with $h_n \to f/q$ uniformly, and then $qh_n \to f$ uniformly with $qh_n \in \mathcal{H}_{0,\varepsilon}$. □

B.2.1 Characteristicness and strict positive definiteness

Universality yields injectivity of the mean embedding on finite signed measures. Applying this injectivity to atomic measures gives strict positive definiteness on distinct nodes. We state both consequences together and prove them separately.

Corollary 3 (Characteristic kernel and SPD)**.** *For every $b_0 > 0$, $\varepsilon_0 > 0$, and every compact $\mathcal{X} \subset \mathbb{R}^d$: (i) k_{b_0,ε_0} is characteristic on $\mathcal{X}$, so the kernel mean embedding $\mu \mapsto \int k_{b_0,\varepsilon_0}(\cdot, \mathbf{x})\, d\mu(\mathbf{x})$ is injective on finite signed Borel measures and the maximum mean discrepancy $\mathrm{MMD}_{k_{b_0,\varepsilon_0}}$ metrizes weak convergence of Borel probability measures on $\mathcal{X}$ (Sriperumbudur et al., 2010, Thm. 23); (ii) k_{b_0,ε_0} is strictly positive definite in the standard kernel-theoretic sense: for any $n \ge 1$ and any n pairwise distinct $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, the Gram matrix $\mathbf{K}$ is strictly positive definite, hence invertible, and the quadratic form $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$ is the squared RKHS norm, inducing a norm on coefficient space by its square root.*

Proof of Corollary 3(i). By Proposition 3, k_{b_0,ε_0} is universal on $\mathcal{X}$. A universal continuous PSD kernel on compact $\mathcal{X}$ is characteristic: if $\mu_k := \int k_{b_0,\varepsilon_0}(\cdot, \mathbf{x})\, d\mu(\mathbf{x}) = 0$ in $\mathcal{H}_{b_0,\varepsilon_0}$ for some finite signed Borel μ, then by the reproducing property $\int f\, d\mu = \langle f, \mu_k \rangle = 0$ for every $f \in \mathcal{H}_{b_0,\varepsilon_0}$.

By density, $\int f\, d\mu = 0$ for every $f \in C(\mathcal{X})$, so $\mu = 0$ by the Riesz representation theorem. On Borel probability measures on $\mathcal{X}$, the metrization of weak convergence by $\mathrm{MMD}_{k_{b_0,\varepsilon_0}}$ follows from Sriperumbudur et al. (2010, Thm. 23). □

Proof of Corollary 3(ii). Let $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ and set $\mu := \sum_{i=1}^n c_i \delta_{\mathbf{x}_i}$, the finite signed Borel measure on $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with atomic masses c_i. By bilinearity of the kernel mean embedding,

$$\mathbf{c}^\top \mathbf{K} \mathbf{c} \;=\; \sum_{i,j} c_i c_j \, k_{b_0,\varepsilon_0}(\mathbf{x}_i, \mathbf{x}_j) \;=\; \Big\| \int k_{b_0,\varepsilon_0}(\cdot, \mathbf{x}) \, d\mu(\mathbf{x}) \Big\|^2_{\mathcal{H}_{b_0,\varepsilon_0}}.$$

If $\mathbf{c}^\top \mathbf{K} \mathbf{c} = 0$, the mean embedding of μ in $\mathcal{H}_{b_0,\varepsilon_0}$ is zero. Since k_{b_0,ε_0} is characteristic on every compact $\mathcal{X} \subset \mathbb{R}^d$ (Corollary 3(i)), the mean embedding is injective on finite signed Borel measures, hence $\mu = 0$. Distinctness of $\{\mathbf{x}_i\}$ then forces $c_i = 0$ for every i. Equivalently, $\mathbf{K} \succ 0$. □

B.3 Section bounds and exact supremum

The first bound below is uniform over bounded inputs and centers. The second fixes the center and optimizes over all inputs, giving an exact global supremum for each section. Neither should be confused with the diagonal value used in the RKHS-ball Rademacher bound.

Proposition 5 (Compact-domain bound and sharp global section bound)**.** *For $\|\mathbf{x}\| \le R$, $\|\mathbf{w}\| \le W$, $b \ge 0$, $\varepsilon > 0$,*

$$0 \;\le\; k_{b,\varepsilon}(\mathbf{w}, \mathbf{x}) \;\le\; \frac{(RW + b)^2}{\varepsilon}.$$

For every fixed $\mathbf{w}$, $b \ge 0$, $\varepsilon > 0$, the section $k_{b,\varepsilon}(\mathbf{w}, \cdot)$ is globally bounded on $\mathbb{R}^d$, with exact supremum

$$\sup_{\mathbf{x} \in \mathbb{R}^d} k_{b,\varepsilon}(\mathbf{w}, \mathbf{x}) = \|\mathbf{w}\|^2 + \frac{(\|\mathbf{w}\|^2 + b)^2}{\varepsilon}. \tag{10}$$

If $(\mathbf{w}, b) \neq (\mathbf{0}, 0)$, equality is attained at $\mathbf{x}^* = \mathbf{w} + \varepsilon \mathbf{w}/(\|\mathbf{w}\|^2 + b)$; when $\mathbf{w} = \mathbf{0}$ and $b > 0$, this gives the maximizer $\mathbf{x} = \mathbf{0}$. For $\mathbf{w} = \mathbf{0}$ and $b = 0$, the section is identically zero.

Proof. Compact-domain bound. By Cauchy–Schwarz, $|\mathbf{w}^\top \mathbf{x}| \le \|\mathbf{w}\| \|\mathbf{x}\| \le WR$, so $|\mathbf{w}^\top \mathbf{x} + b| \le WR + b$. Also $D_\varepsilon(\mathbf{x}, \mathbf{w}) \ge \varepsilon > 0$, so $k_{b,\varepsilon}(\mathbf{w}, \mathbf{x}) \le (WR + b)^2/\varepsilon$. Nonnegativity follows from the squared numerator and positive denominator.

Exact global supremum. Put $\mathbf{y} = \mathbf{x} - \mathbf{w}$ and $c = \|\mathbf{w}\|^2 + b$. Cauchy–Schwarz in $\mathbb{R}^{d+1}$ gives

$$(\mathbf{w}^\top \mathbf{y} + c)^2 = \big\langle (\mathbf{w}, c/\sqrt{\varepsilon}), (\mathbf{y}, \sqrt{\varepsilon}) \big\rangle^2 \le \left(\|\mathbf{w}\|^2 + \frac{c^2}{\varepsilon} \right) (\|\mathbf{y}\|^2 + \varepsilon).$$

Division by $\|\mathbf{y}\|^2 + \varepsilon$ gives the upper bound in (10). To show that it is attained, first suppose $c > 0$. Equality holds when $(\mathbf{y}, \sqrt{\varepsilon}) = (\varepsilon/c)(\mathbf{w}, c/\sqrt{\varepsilon})$, namely at $\mathbf{y} = \varepsilon \mathbf{w}/c$. If $c = 0$, nonnegativity of b forces $\mathbf{w} = \mathbf{0}$ and $b = 0$, and the section is zero. □

Diagonal value and the capacity bound. The diagonal value $k_{b,\varepsilon}(\mathbf{x}, \mathbf{x}) = (\|\mathbf{x}\|^2 + b)^2/\varepsilon$ used in the Rademacher proof is obtained by setting $\mathbf{w} = \mathbf{x}$ in the kernel definition, not by maximizing a fixed-center section.

Classical scalar activations such as ReLU and GELU are unbounded along positive-projection directions. Their dot-product parameterization does not by itself supply a PSD kernel on the shared input–parameter domain, so this particular kernel-diagonal Rademacher calculation does not automatically apply to them.

C What Alignment Changes: Proofs and Extensions

The alignment factor has two effects studied separately here. Its directional trace distinguishes Yat sections from the IMQ RKHS; its projection dependence determines activation covariance under Gaussian and fixed-norm inputs. We prove these results in turn, then give uniform exterior-shell separation bounds under explicit center and coefficient constraints.

C.1 Directional structure and quantitative separation

We first show that every IMQ-RKHS function vanishes at infinity. The Yat channel decomposition and section traces then identify functions outside that RKHS. The trace is defined on finite spans: the nonclosability result below explains why it cannot simply be extended to the whole ambient RKHS.

Proof of Lemma 1. Every finite-span element $f_m(\mathbf{x}) = \sum_{j=1}^m a_j h_\varepsilon(\mathbf{x}, \mathbf{w}_j)$ vanishes at infinity since each atom decays as $O(\|\mathbf{x}\|^{-2})$. The finite span is dense in $\mathcal{H}_{h_\varepsilon}$; if $f_m \to f$ in RKHS norm, the reproducing property gives

$$\sup_{\mathbf{x}\in\mathbb{R}^d} |f_m(\mathbf{x}) - f(\mathbf{x})| \le \sup_{\mathbf{x}} \sqrt{h_\varepsilon(\mathbf{x},\mathbf{x})}\, \|f_m - f\|_{\mathcal{H}_{h_\varepsilon}} = \varepsilon^{-1/2}\|f_m - f\|_{\mathcal{H}_{h_\varepsilon}},$$

so $f_m \to f$ uniformly. Since $C_0(\mathbb{R}^d)$ is closed under uniform limits, $f \in C_0(\mathbb{R}^d)$. □

Theorem 8 (Yat RKHS channel decomposition)**.** *Write $k_{b,\varepsilon} = k_0 + k_1 + k_2$ where $k_0 := b^2 h_\varepsilon$, $k_1 := 2b(\mathbf{x}^\top\mathbf{w})h_\varepsilon$, $k_2 := (\mathbf{x}^\top\mathbf{w})^2 h_\varepsilon$. By the RKHS sum rule (Paulsen & Raghupathi, 2016; Berlinet & Thomas-Agnan, 2004),*

$$\mathcal{H}_{b,\varepsilon} = \mathcal{H}_{k_0} + \mathcal{H}_{k_1} + \mathcal{H}_{k_2},$$

with squared infimal-convolution norm

$$\|f\|^2_{\mathcal{H}_{b,\varepsilon}} = \inf_{\substack{f_0+f_1+f_2=f \\ f_i\in\mathcal{H}_{k_i}}} \big(\|f_0\|^2_{\mathcal{H}_{k_0}} + \|f_1\|^2_{\mathcal{H}_{k_1}} + \|f_2\|^2_{\mathcal{H}_{k_2}}\big).$$

Proof of Theorem 8. The RKHS sum rule states that $\mathcal{H}_{k_1+k_2} = \mathcal{H}_{k_1} + \mathcal{H}_{k_2}$ with the infimal-convolution norm (Paulsen & Raghupathi, 2016); induction extends the rule to three summands. It remains to check that each channel is PSD.

The radial channel is PSD: $k_0 = b^2 h_\varepsilon$ since h_ε is PSD and $b^2 \ge 0$. The linear-alignment channel $k_1 = 2b(\mathbf{x}^\top\mathbf{w})h_\varepsilon$ is PSD because the dot-product kernel $\mathbf{x}^\top\mathbf{w}$ is PSD ($\sum_{ij} c_i c_j \mathbf{x}_i^\top\mathbf{x}_j = \|\sum_i c_i\mathbf{x}_i\|^2 \ge 0$), h_ε is PSD, their Schur product is PSD by the Schur product theorem, and $2b \ge 0$ (the channel vanishes at $b = 0$). Finally, the quadratic-alignment channel $k_2 = (\mathbf{x}^\top\mathbf{w})^2 h_\varepsilon$ is PSD by the Schur product theorem applied to $(\mathbf{x}^\top\mathbf{w})^2$ and h_ε, both PSD. The RKHS sum rule therefore applies. □

For $b > 0$, the *radial IMQ channel* $k_0 = b^2 h_\varepsilon$ has $\mathcal{H}_{k_0} = \mathcal{H}_{h_\varepsilon}$ as sets, with rescaled norm $\|f\|_{\mathcal{H}_{k_0}} = b^{-1}\|f\|_{\mathcal{H}_{h_\varepsilon}}$ by the constant-rescaling identity Paulsen & Raghupathi 2016. The *linear-alignment IMQ channel* $k_1 = 2b(\mathbf{x}^\top\mathbf{w})h_\varepsilon$ also vanishes at $b = 0$. The *quadratic-alignment IMQ channel* $k_2 = (\mathbf{x}^\top\mathbf{w})^2 h_\varepsilon$ remains for every $b \ge 0$ and carries the far-field trace.

Under the zero-kernel convention fixed above, both k_0 and k_1 vanish at $b = 0$: the RKHS no longer contains the full IMQ RKHS and loses the constant coordinate, which is the degeneracy responsible for failure of fixed-kernel universality on domains containing the origin. Equivalently, $k_{0,\varepsilon}(\mathbf{0},\mathbf{0}) = 0$ forces every $f \in \mathcal{H}_{0,\varepsilon}$ to satisfy $f(\mathbf{0}) = 0$. The channel decomposition and the vanishing diagonal thus identify the same endpoint degeneracy.

Proof of Proposition 1. Yat limit. For (3), write

$$g_\varepsilon(r\mathbf{u};\mathbf{w},b) \;=\; \frac{(r\,\mathbf{u}^\top\mathbf{w} + b)^2}{r^2 - 2r\,\mathbf{u}^\top\mathbf{w} + \|\mathbf{w}\|^2 + \varepsilon} \;=\; \frac{(\mathbf{u}^\top\mathbf{w} + b/r)^2}{1 - 2(\mathbf{u}^\top\mathbf{w})/r + (\|\mathbf{w}\|^2+\varepsilon)/r^2} \;\xrightarrow[r\to\infty]{}\; (\mathbf{u}^\top\mathbf{w})^2,$$

uniformly in $\mathbf{u} \in \mathbb{S}^{d-1}$.

IMQ decay. For each fixed center $\mathbf{w}_i$, $|k^\varepsilon_{\mathrm{IMQ}}(r\mathbf{u},\mathbf{w}_i)| \le (r^2 - 2r\|\mathbf{w}_i\| + \|\mathbf{w}_i\|^2 + \varepsilon)^{-1} = O(r^{-2})$ uniformly on the sphere. Hence every finite combination $F = \sum_i c_i k^\varepsilon_{\mathrm{IMQ}}(\cdot,\mathbf{w}_i)$ satisfies $T_\infty F \equiv 0$.

Range of the trace. The image of T_∞ contains every rank-one quadratic form $(\mathbf{u}^\top \mathbf{w})^2 = \mathbf{u}^\top(\mathbf{w}\mathbf{w}^\top)\mathbf{u}$. The symmetric rank-one tensors $\mathbf{w}\mathbf{w}^\top$ span $\mathrm{Sym}(d)$: the diagonal generators are $\mathbf{e}_i\mathbf{e}_i^\top$, and the off-diagonal generators follow from $\frac{1}{2}((\mathbf{e}_i+\mathbf{e}_j)(\mathbf{e}_i+\mathbf{e}_j)^\top - \mathbf{e}_i\mathbf{e}_i^\top - \mathbf{e}_j\mathbf{e}_j^\top)$. Therefore $\mathrm{Im}(T_\infty) = \mathcal{Q}_2(\mathbb{S}^{d-1})$. □

The trace formula does not imply continuity in the ambient RKHS norm. The following construction keeps the trace fixed while the RKHS norm tends to zero.

Proposition 6 (The finite-span trace is not closable in the ambient RKHS)**.** *Fix $b \geq 0$, $\varepsilon > 0$, and $d \geq 1$. On the finite section span of $k_{b,\varepsilon}$, the trace T_∞ is not closable as an operator from $\mathcal{H}_{b,\varepsilon}$ to $C(\mathbb{S}^{d-1})$ with the uniform norm. In particular, it has no bounded extension to the entire ambient RKHS.*

Proof. A fixed trace with vanishing RKHS norm. Fix a unit vector $\mathbf{u}$ and a spacing $L > 0$. Put $\mathbf{w}_j = jL\mathbf{u}$ and

$$f_N = \frac{1}{N}\sum_{j=1}^{N} \frac{k_{b,\varepsilon}(\mathbf{w}_j, \cdot)}{(jL)^2}.$$

The section-trace formula gives $T_\infty f_N(\mathbf{v}) = (\mathbf{u}^\top \mathbf{v})^2 =: q(\mathbf{v})$ for every N, with $\|q\|_\infty = 1$. The reproducing property gives

$$\|f_N\|^2_{\mathcal{H}_{b,\varepsilon}} = \frac{1}{N^2}\sum_{i,j=1}^{N} \frac{(1+b/(ijL^2))^2}{(i-j)^2L^2+\varepsilon} \leq \frac{(1+b/L^2)^2}{N}\left(\frac{1}{\varepsilon}+\frac{4}{L^2}\right).$$

Indeed, each row of reciprocal denominators sums to at most $\varepsilon^{-1} + 2L^{-2}\sum_{h\geq 1} h^{-2} \leq \varepsilon^{-1} + 4L^{-2}$, by the integral bound on that series. Thus $f_N \to 0$ in the RKHS while $T_\infty f_N = q \neq 0$ uniformly, which violates the defining sequential criterion for closability.

Behavior on bounded domains. On every radius-R ball the reproducing inequality gives $\sup_{\|\mathbf{x}\|\leq R} |f_N(\mathbf{x})| \leq (R^2+b)\|f_N\|_{\mathcal{H}_{b,\varepsilon}}/\sqrt{\varepsilon} \to 0$. The construction requires unrestricted centers; it makes no assertion about bounded-center or fixed-dictionary classes. □

Proof of Corollary 1. The inclusion $\mathcal{H}_{h_\varepsilon} \subseteq \mathcal{H}_{b,\varepsilon}$ is Theorem 7. For strictness, choose $\mathbf{w} \neq \mathbf{0}$ and any $\mathbf{u}$ with $\mathbf{u}^\top\mathbf{w} \neq 0$: the far-field trace gives $\lim_{r\to\infty} k_{b,\varepsilon}(\mathbf{w}, r\mathbf{u}) = (\mathbf{u}^\top\mathbf{w})^2 \neq 0$, so $k_{b,\varepsilon}(\mathbf{w},\cdot) \notin C_0(\mathbb{R}^d)$ while, by Lemma 1, $\mathcal{H}_{h_\varepsilon} \subset C_0(\mathbb{R}^d)$. Hence $k_{b,\varepsilon}(\mathbf{w},\cdot) \in \mathcal{H}_{b,\varepsilon} \setminus \mathcal{H}_{h_\varepsilon}$.

For non-reversibility, suppose $k_{b,\varepsilon} \preceq C\, h_\varepsilon$. Then Aronszajn inclusion would give $\mathcal{H}_{b,\varepsilon} \subseteq \mathcal{H}_{h_\varepsilon} \subset C_0(\mathbb{R}^d)$, contradicting the existence of Yat sections not vanishing at infinity. □

Corollary 4 (Concrete large-radius regression gap)**.** *Fix $\mathbf{w} \neq \mathbf{0}$, $b \in \mathbb{R}$, $\varepsilon > 0$, and choose $\mathbf{u} \in \mathbb{S}^{d-1}$ with $a := (\mathbf{u}^\top\mathbf{w})^2 > 0$. Let $f(\mathbf{x}) = g_\varepsilon(\mathbf{x}; \mathbf{w}, b)$. Then for every finite IMQ combination $F \in F_{\mathrm{IMQ},\varepsilon}$,*

$$|f(r\mathbf{u}) - F(r\mathbf{u})| \xrightarrow[r\to\infty]{} a.$$

Hence for every $\eta \in (0, a)$ there exists R_η such that $\sup_{r\geq R_\eta} |f(r\mathbf{u}) - F(r\mathbf{u})| \geq a - \eta$. In particular, a single Yat atom carries an $O(1)$ directional signal on a sufficiently large exterior ray that no finite IMQ expansion can uniformly match there.

Proof of Corollary 4. Proposition 1 gives $f(r\mathbf{u}) \to a$ and $F(r\mathbf{u}) \to 0$, hence $f(r\mathbf{u}) - F(r\mathbf{u}) \to a$. Given $\eta \in (0, a)$, choose R_η so that for all $r \geq R_\eta$: $|f(r\mathbf{u}) - a| < \eta/2$ and $|F(r\mathbf{u})| < \eta/2$. Then $|f(r\mathbf{u}) - F(r\mathbf{u})| \geq a - \eta$, and the supremum over $r \geq R_\eta$ preserves the bound. □

The same far-field calculation can be made for derivatives with respect to the center. The next result concerns individual sections; it does not assert an optimization guarantee for a trained network.

Proposition 7 (Far-field gradient asymptotics)**.** *Fix $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, $b \geq 0$, $\varepsilon > 0$, and a unit direction $\mathbf{u} \in \mathbb{S}^{d-1}$ with $\mathbf{u}^\top \mathbf{w} \neq 0$. Along the ray $\mathbf{x} = r\mathbf{u}$, the center-gradients of the IMQ section $h_\varepsilon(\mathbf{x}, \mathbf{w})$ and the Yat section $k_{b,\varepsilon}(\mathbf{w}, \mathbf{x})$ satisfy*

$$\left\|\nabla_{\mathbf{w}} h_\varepsilon(r\mathbf{u}, \mathbf{w})\right\| = \Theta(r^{-3}), \qquad \lim_{r\to\infty} \nabla_{\mathbf{w}} k_{b,\varepsilon}(\mathbf{w}, r\mathbf{u}) = 2(\mathbf{u}^\top \mathbf{w})\, \mathbf{u}.$$

Remark 1. *The proposition records an asymptotic difference in the center-gradient of a single Yat atom versus a single IMQ atom at large $\|\mathbf{x}\|$, evaluated in isolation. It is a statement about the kernel itself, not about gradients of trained networks: in a finite expansion the per-atom gradients combine through the readout coefficients $\boldsymbol{\alpha}$, and far-field signal can be informative or uninformative depending on where the data actually live.*

Proof of Proposition 7. For the IMQ atom, $\nabla_{\mathbf{w}} h_\varepsilon = 2(\mathbf{x}-\mathbf{w})/(\|\mathbf{x}-\mathbf{w}\|^2+\varepsilon)^2$. With $\mathbf{x} = r\mathbf{u}$, the numerator $\|r\mathbf{u} - \mathbf{w}\| = r + O(1)$ and the denominator $(r^2 - 2r\,\mathbf{u}^\top \mathbf{w} + \|\mathbf{w}\|^2 + \varepsilon)^2 = r^4 + O(r^3)$, so $\|\nabla_{\mathbf{w}} h_\varepsilon\| = 2r^{-3}(1 + O(r^{-1})) = \Theta(r^{-3})$.

For the Yat atom, set $N(\mathbf{x}) := (\mathbf{w}^\top \mathbf{x} + b)^2$ and $D(\mathbf{x}) := \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$. The quotient rule gives

$$\nabla_{\mathbf{w}} k_{b,\varepsilon}(\mathbf{w}, \mathbf{x}) = \frac{2(\mathbf{w}^\top \mathbf{x} + b)\, \mathbf{x}}{D(\mathbf{x})} + \frac{2(\mathbf{w}^\top \mathbf{x} + b)^2 (\mathbf{x} - \mathbf{w})}{D(\mathbf{x})^2}.$$

Substituting $\mathbf{x} = r\mathbf{u}$, the first term equals $2(r\mathbf{u}^\top \mathbf{w} + b)\, r\mathbf{u}/D(r\mathbf{u})$. The leading-order behavior of numerator and denominator is $2r^2(\mathbf{u}^\top \mathbf{w})\mathbf{u} + O(r)$ over $r^2 + O(r)$, giving the limit $2(\mathbf{u}^\top \mathbf{w})\mathbf{u}$. The second term has numerator $2(r\mathbf{u}^\top \mathbf{w} + b)^2 (r\mathbf{u} - \mathbf{w}) = O(r^3)$ over denominator $D(r\mathbf{u})^2 = r^4 + O(r^3)$, giving $O(r^{-1}) \to 0$. Summing yields the stated limit. □

C.2 Gaussian and fixed-norm activation covariance

We now turn from deterministic far-field behavior to distribution-dependent activation covariance. The input distribution determines the leading radial covariance. Under isotropic Gaussian input, IMQ units share the fluctuation of $\|\mathbf{x}\|^2$ and their leading covariance is rank one, while the Yat numerator retains order-one relative variation. On the fixed-norm sphere, the shared radial fluctuation vanishes and center-specific projections give both kernels full-rank limiting covariance at orthogonal centers. Related high-dimensional linearization and low-degree phenomena are well documented: El Karoui (2010) shows that broad classes of radial kernel random matrices become essentially linear on isotropic high-dimensional inputs, while Ghorbani et al. (2019) derives low-degree approximation limits for rotationally invariant random-feature and neural-tangent models. The proposition below gives a different, activation-level statement for the present IMQ/Yat pair: under isotropic Gaussian inputs, the Yat numerator $(\mathbf{x}^\top \mathbf{w} + b)^2$ depends on a one-dimensional projection and retains nonvanishing relative variation as the dimension grows.

Proposition 8 (High-dimensional relative variation)**.** *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, let $\mathbf{w} \in \mathbb{R}^d$ have unit norm $\|\mathbf{w}\| = 1$, and fix $b \geq 0$, $\varepsilon > 0$. For a positive random variable Y, write $\mathrm{CV}[Y] := \sqrt{\mathrm{Var}[Y]}/\mathbb{E}[Y]$. As $d \to \infty$,*

$$\sqrt{d}\,\mathrm{CV}[h_\varepsilon(\mathbf{x}, \mathbf{w})] \longrightarrow \sqrt{2}, \qquad \mathrm{CV}[k_{b,\varepsilon}(\mathbf{w}, \mathbf{x})] \longrightarrow \frac{\sqrt{2 + 4b^2}}{1 + b^2}.$$

Proof. By rotational invariance of the standard Gaussian, take $\mathbf{w} = \mathbf{e}_1$. Then $\mathbf{w}^\top \mathbf{x} = x_1 \sim \mathcal{N}(0, 1)$, and the regularized distance $D := \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon = (x_1 - 1)^2 + \sum_{i=2}^d x_i^2 + \varepsilon$ satisfies $D - \varepsilon \sim \chi^2_d(\lambda = 1)$ (non-central chi-squared, d degrees of freedom, non-centrality $\|\mathbf{w}\|^2 = 1$). Standard moment formulas give $\mu_D := \mathbb{E}[D] = d + 1 + \varepsilon$ and $\sigma_D^2 := \mathrm{Var}[D] = 2d + 4$.

Inverse-distance concentration. The non-central chi-square law gives

$$\frac{D}{d} \longrightarrow 1 \quad \text{and} \quad \frac{D - \mu_D}{\sqrt{d}} \Longrightarrow \mathcal{N}(0, 2).$$

The required moment convergence follows from an explicit event split. On $\mathcal{E}_d := \{D \geq d/2\}$, $d/D \leq 2$. On $\mathcal{E}_d^c$, the deterministic floor $D \geq \varepsilon$ and a non-central chi-square lower-tail

bound give, for every fixed $q < \infty$, $\mathbb{E}[(d/D)^q \mathbf{1}_{\mathcal{E}_d^c}] \le (d/\varepsilon)^q e^{-cd} = o(1)$. Since $D/d \to 1$ in probability, this proves $d/D \to 1$ in L^q for every fixed q. Moreover, the standardized non-central chi-square variables $(D - \mu_D)/\sqrt{d}$ have uniformly bounded moments of every fixed order. These two facts make the squared variables in the following delta-method limit uniformly integrable:

$$d^{3/2}\left(\frac{1}{D} - \frac{1}{\mu_D}\right) = -\frac{D - \mu_D}{\sqrt{d}} \frac{d^2}{D\mu_D} \Longrightarrow \mathcal{N}(0, 2),$$

with convergence of second moments. Consequently,

$$\mathbb{E}[h_\varepsilon] = d^{-1}(1 + o(1)), \qquad \mathrm{Var}[h_\varepsilon] = 2d^{-3}(1 + o(1)),$$

and hence $\sqrt{d}\,\mathrm{CV}[h_\varepsilon] \to \sqrt{2}$.

Yat unit. Let $N := (x_1 + b)^2$. Then $k_{b,\varepsilon} = N/D$, $\mathbb{E}N = 1 + b^2$, and $\mathrm{Var}(N) = 2 + 4b^2 > 0$. The L^4 convergence $d/D \to 1$ and Hölder's inequality give $\mathbb{E}[N^2(d/D - 1)^2] \le (\mathbb{E}N^4)^{1/2}(\mathbb{E}|d/D - 1|^4)^{1/2} \to 0$. Hence

$$\frac{dN}{D} \longrightarrow N \qquad \text{in } L^2.$$

No independence between N and D is required. Taking first and second moments,

$$d\,\mathbb{E}[k_{b,\varepsilon}] \longrightarrow 1 + b^2, \qquad d^2\,\mathrm{Var}[k_{b,\varepsilon}] \longrightarrow 2 + 4b^2.$$

Therefore $\mathrm{CV}[k_{b,\varepsilon}] \to \sqrt{2 + 4b^2}/(1 + b^2)$. □

The single-unit calculation extends to joint activations. Under Gaussian inputs, the leading IMQ fluctuation is common to all centers, whereas the Yat limit retains their projection correlations.

Proof of Theorem 1(i). Write $Z_{j,d} := \mathbf{w}_{j,d}^\top \mathbf{x}_d$, $S_d := \|\mathbf{x}_d\|^2$, and $D_{j,d} := S_d - 2Z_{j,d} + 1 + \varepsilon$. The vector $\mathbf{Z}_d = (Z_{1,d}, \ldots, Z_{m,d})$ is Gaussian with covariance $\mathbf{C}_d$ and therefore converges in distribution, with convergence of every fixed joint moment, to $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$.

Radial covariance. Let $\mu_d := d + 1 + \varepsilon$ and $G_d := (S_d - d)/\sqrt{d}$. The same lower-tail event split used above, applied to the fixed collection $D_{1,d}, \ldots, D_{m,d}$, gives $d/D_{j,d} \to 1$ in every fixed L^q. Since

$$\frac{D_{j,d} - \mu_d}{\sqrt{d}} = G_d - \frac{2Z_{j,d}}{\sqrt{d}},$$

the second term vanishes in every fixed L^q, while $G_d \Rightarrow G \sim \mathcal{N}(0, 2)$ with convergence of moments. Since m is fixed, $d/D_{j,d} \to 1$ in probability holds jointly over $j = 1, \ldots, m$. Slutsky's theorem therefore gives the joint distributional convergence

$$d^{3/2}\left(\frac{1}{D_{j,d}} - \frac{1}{\mu_d}\right) = -\frac{D_{j,d} - \mu_d}{\sqrt{d}} \frac{d^2}{D_{j,d}\mu_d} \Longrightarrow -G \qquad (j = 1, \ldots, m).$$

This convergence also passes to cross-moments. Indeed, the lower-tail event split gives uniform bounds on $d/D_{j,d}$ in every fixed L^q, while $(D_{j,d} - \mu_d)/\sqrt{d}$ has uniformly bounded moments of every fixed order. Hölder's inequality therefore bounds the displayed variables uniformly in $L^{2+\eta}$ for some $\eta > 0$. Their products are uniformly integrable, so joint distributional convergence implies convergence of first and cross-moments. Since $\mathbb{E}G = 0$, centering therefore yields $d^3\,\mathrm{Cov}(D_{i,d}^{-1}, D_{j,d}^{-1}) \to \mathbb{E}G^2 = 2$ for every i, j.

Yat covariance. Set $Q_{j,d} := (Z_{j,d} + b)^2$. Joint convergence of $\mathbf{Z}_d$, the continuous mapping theorem, and the joint convergence $d/D_{j,d} \to 1$ in probability give

$$d\,k_{b,\varepsilon}(\mathbf{w}_{j,d}, \mathbf{x}_d) = Q_{j,d}\frac{d}{D_{j,d}} \Longrightarrow (Z_j + b)^2 \qquad \text{jointly in distribution.}$$

The fixed Gaussian moments of $Q_{j,d}$, the uniform L^q bounds for $d/D_{j,d}$, and Hölder's inequality bound the left-hand variables uniformly in $L^{2+\eta}$ for some $\eta > 0$. Their products

are therefore uniformly integrable, so first and cross-moments converge to those of $((Z_j + b)^2)_{j=1}^m$. For jointly centered Gaussian Z_i, Z_j with $\mathbb{E}[Z_i Z_j] = C_{ij}$, Isserlis' formula gives

$$\mathrm{Cov}\big((Z_i + b)^2, (Z_j + b)^2\big) = 2C_{ij}^2 + 4b^2 C_{ij}.$$

Taking the resulting covariance limits proves the second limit in (4).

Matrix limits and empirical covariance. Because m is fixed, entrywise convergence is equivalent to convergence in operator norm. When $\mathbf{C} = \mathbf{I}_m$, the two displayed eigenspectra follow directly. Finally, the strong law applied entrywise to the first and second activation moments gives almost-sure convergence of each empirical covariance at fixed d. □

Spherical normalization removes the shared norm fluctuation. The proof below therefore expands around the deterministic squared radius d, leaving the center-specific projections as the leading IMQ variation.

Proof of Theorem 1(ii). Write $Z_{j,d} = \mathbf{w}_{j,d}^\top \mathbf{x}_d$ and $a = 1 + \varepsilon$. The finite projection vector $\mathbf{Z}_d = (Z_{j,d})_{j=1}^m$ converges in distribution, and its joint moments of every fixed order converge to those of $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, by the spherical moment formula. Since $\|\mathbf{x}_d\|^2 = d$ exactly,

$$D_{j,d} := \|\mathbf{x}_d - \mathbf{w}_{j,d}\|^2 + \varepsilon = d + a - 2Z_{j,d}.$$

The bound $|Z_{j,d}| \le \sqrt{d}$ gives $D_{j,d} \ge (\sqrt{d} - 1)^2 + \varepsilon$. Expanding the multiplier and using the spherical moments gives

$$\left\| d^2 \left(H_{j,d} - \frac{1}{d} \right) - (2Z_{j,d} - a) \right\|_{L^2} = \left\| (2Z_{j,d} - a) \left(\frac{d}{D_{j,d}} - 1 \right) \right\|_{L^2} \longrightarrow 0.$$

Covariance is unchanged by subtracting $1/d$. Joint moment convergence of $\mathbf{Z}_d$ now gives the first limit.

For the Yat coordinates,

$$\left\| dY_{j,d} - (Z_{j,d} + b)^2 \right\|_{L^2} = \left\| (Z_{j,d} + b)^2 \left(\frac{d}{D_{j,d}} - 1 \right) \right\|_{L^2} \longrightarrow 0.$$

All variables in each norm live on the d-dimensional spherical probability space; the Gaussian vector enters only after passing the joint moments of $\mathbf{Z}_d$ to their limits. For jointly centered Gaussian G_i, G_j with covariance C_{ij}, $\mathrm{Cov}((G_i + b)^2, (G_j + b)^2) = 2C_{ij}^2 + 4b^2 C_{ij}$. This proves the second limit. Since m is fixed, entrywise convergence is equivalent to operator-norm convergence. □

Rank through quadratic lifting. The covariance formulas reduce the rank question to the span of the lifted center features.

Proof of Corollary 2. Tensor inner products give $\langle \mathbf{v}_i \otimes \mathbf{v}_i, \mathbf{v}_j \otimes \mathbf{v}_j \rangle = C_{ij}^2$. Hence the lifted Gram matrix is $2(\mathbf{C} \circ \mathbf{C}) + 4b^2\mathbf{C}$. The rank of a Gram matrix equals the dimension of the span of its features. The dimension bounds follow because $\mathbf{v} \otimes \mathbf{v}$ lies in $\mathrm{Sym}(r)$; the IMQ claim follows from its limit $4\mathbf{C}$. □

To verify the main-text example, let the three limiting centers be $\mathbf{e}_1, \mathbf{e}_2, (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$. Then $\operatorname{rank} \mathbf{C} = 2$ while $\det(\mathbf{C} \circ \mathbf{C}) = 1/2 > 0$. Thus at $b = 0$ the spherical IMQ limit has rank two and the Yat limit has rank three. For $b > 0$, the added term $4b^2\mathbf{C}$ is PSD, so the Yat limit remains positive definite and has rank three.

Remark 2. *The two cases of Theorem 1 separate two mechanisms. Gaussian inputs contain an order-$\sqrt{d}$ squared-norm fluctuation shared by every center, which dominates the centered IMQ covariance. Spherical normalization removes that common term, exposing the center-specific projection covariance $4\mathbf{C}$. The Yat numerator already depends at leading order on those projections, so its covariance limit is the same in both models. These two distributions provide analytically exact reference regimes for studying normalized neural representations.*

The preceding limits keep the number of centers fixed. The following finite-dimensional estimates instead compare covariance matrices directly with the current center Gram matrix and make the dependence on the number of centers explicit.

C.2.1 Explicit finite-dimensional covariance control

Write $\mathcal{L}_b(\mathbf{C}) = 2(\mathbf{C} \circ \mathbf{C}) + 4b^2\mathbf{C}$. The following conservative bounds are uniform over unit-center configurations; they do not require a limiting Gram matrix or a fixed number of centers. For fixed kernel parameters, the operator-norm errors scale as $m/\sqrt{d}$ for Gaussian inputs and m/d for spherical inputs. The explicit constants below make these population bounds usable without an asymptotic limit.

Proposition 9 (Finite-dimensional covariance remainders)**.** *Let $d \geq 34$, $m \geq 1$, $b \geq 0$, and $\varepsilon > 0$. For any m unit centers in $\mathbb{R}^d$, let $\mathbf{C}_d$ be their Gram matrix and use the activation vectors $\mathbf{H}_d, \mathbf{Y}_d$ of Theorem 1. The constants below separate numerator moment bounds (Q_2, Q_4), Gaussian denominator control (M_d, u_d), the spherical denominator lower bound (ℓ_d), and activation approximation errors (ρ). Define*

$$a = 1 + \varepsilon, \quad z_4 = 3^{1/4}, \quad z_8 = 105^{1/8}, \quad Q_2 = (z_4 + b)^2, \quad Q_4 = (z_8 + b)^2,$$
$$M_d = (1680d^4 + 76160d^3 + 467712d^2 + 645120d)^{1/8},$$
$$u_d = \frac{2(M_d + 2z_8 + a)}{d}, \qquad \ell_d = d + a - 2\sqrt{d} > 0,$$
$$\rho_{Y,G} = Q_4 u_d, \qquad \rho_{H,G} = (12 + 48/d)^{1/4} u_d + \frac{2(2z_4 + a)}{\sqrt{d}},$$
$$\rho_{Y,S} = \frac{Q_4(2z_4 + a)}{\ell_d}, \qquad \rho_{H,S} = \frac{(2z_4 + a)^2}{\ell_d}.$$

Under Gaussian inputs,

$$\|d^3 \operatorname{Cov}(\mathbf{H}_d) - 2\mathbf{1}\mathbf{1}^\top\|_{\mathrm{op}} \leq E_{H,G} := m(2\sqrt{2}\,\rho_{H,G} + \rho_{H,G}^2), \tag{11}$$
$$\|d^2 \operatorname{Cov}(\mathbf{Y}_d) - \mathcal{L}_b(\mathbf{C}_d)\|_{\mathrm{op}} \leq E_{Y,G} := m(2Q_2\rho_{Y,G} + \rho_{Y,G}^2). \tag{12}$$

Under uniform inputs on $\sqrt{d}\,\mathbb{S}^{d-1}$,

$$\|d^4 \operatorname{Cov}(\mathbf{H}_d) - 4\mathbf{C}_d\|_{\mathrm{op}} \leq E_{H,S} := m(4\rho_{H,S} + \rho_{H,S}^2), \tag{13}$$
$$\|d^2 \operatorname{Cov}(\mathbf{Y}_d) - \mathcal{L}_b(\mathbf{C}_d)\|_{\mathrm{op}} \leq E_{Y,S} := m\left(2Q_2\rho_{Y,S} + \rho_{Y,S}^2 + \frac{6}{d+2}\right). \tag{14}$$

For fixed b, ε, the Gaussian errors are $O(m/\sqrt{d})$ and the spherical errors are $O(m/d)$. These are sufficient, not sharp, rates; $d \geq 34$ is a proof threshold, not a failure boundary.

Proof. We first record a covariance perturbation bound on a common probability space. If $\|A_j - B_j\|_{L^2} \leq \rho$ and $\|B_j - \mathbb{E}B_j\|_{L^2} \leq V$ for every j, then

$$\|\operatorname{Cov}(\mathbf{A}) - \operatorname{Cov}(\mathbf{B})\|_{\mathrm{op}} \leq m(2V\rho + \rho^2). \tag{15}$$

Centering is an L^2 contraction. Expanding each covariance difference into two cross terms and one residual term, Cauchy–Schwarz bounds its absolute value by $2V\rho + \rho^2$. The operator norm is at most m times the largest absolute entry.

Gaussian inverse moments. For a unit center, write $Z = \mathbf{w}^\top\mathbf{x}$, $S = \|\mathbf{x}\|^2$, $D = S - 2Z + a$, and $G = (S - d)/\sqrt{d}$. After rotation, $D = (Z - 1)^2 + T + \varepsilon$ with $T \sim \chi^2_{d-1}$. For an integer $q \geq 1$, integration of the chi-square density gives

$$\mathbb{E}T^{-q} = \prod_{s=1}^{q}(d - 1 - 2s)^{-1} \quad (d > 2q + 1).$$

For $q = 8$ and $d \geq 34$, each factor satisfies $d/(d - 1 - 2s) \leq 2$, hence $\|d/D\|_{L^8} \leq 2$. The moment-generating function $\mathbb{E}e^{t(S-d)} = e^{-dt}(1 - 2t)^{-d/2}$ gives $\mathbb{E}(S - d)^4 = 12d^2 + 48d$ and $\mathbb{E}(S - d)^8 = M_d^8$ by coefficient extraction. Since $\|Z\|_{L^4} = z_4$ and $\|Z\|_{L^8} = z_8$, Hölder and Minkowski give

$$\left\|\frac{d}{D} - 1\right\|_{L^4} = \left\|\frac{S - d - 2Z + a}{d}\,\frac{d}{D}\right\|_{L^4} \leq u_d.$$

For $Q = (Z+b)^2$, $\|Q\|_{L^2} \le Q_2$ and $\|Q\|_{L^4} \le Q_4$, so $\|dY - Q\|_{L^2} \le Q_4 u_d = \rho_{Y,G}$. The vector with coordinates $Q_j = (Z_j + b)^2$, formed from the jointly Gaussian projections Z_j, has covariance $\mathcal{L}_b(\mathbf{C}_d)$ exactly. Applying (15) with $V = Q_2$ proves (12). For the IMQ coordinates, subtracting deterministic constants does not affect covariance, and the exact identity

$$d^{3/2}(H - d^{-1}) = \left(-G + \frac{2Z - a}{\sqrt{d}}\right)\frac{d}{D}$$

implies $\|d^{3/2}(H - d^{-1}) + G\|_{L^2} \le \rho_{H,G}$. Use $\|G\|_{L^4} = (12 + 48/d)^{1/4}$ and $\|d/D\|_{L^4} \le 2$ to obtain this inequality. The comparison vector has every coordinate equal to $-G$ and covariance $2\mathbf{1}\mathbf{1}^\top$; its coordinate standard deviation is $\sqrt{2}$. Equation (15) proves (11).

Spherical inputs. Here $D = d + a - 2Z \ge \ell_d$ deterministically. The spherical moment formula

$$\mathbb{E}Z^{2p} = (2p-1)!! \prod_{s=0}^{p-1} \frac{d}{d+2s} \le (2p-1)!!$$

follows, for example, by writing a standard Gaussian vector as its independent radius times a uniform unit direction and comparing even projection moments. Thus the same constants z_4, z_8, Q_2, Q_4 remain valid. The exact residual identities yield

$$\|dY - Q\|_{L^2} = \left\|\frac{Q(2Z-a)}{D}\right\|_{L^2} \le \rho_{Y,S},$$
$$\|d^2(H - d^{-1}) - (2Z - a)\|_{L^2} = \left\|\frac{(2Z-a)^2}{D}\right\|_{L^2} \le \rho_{H,S}.$$

The comparison IMQ vector $2\mathbf{Z} - a\mathbf{1}$ has covariance $4\mathbf{C}_d$ and coordinate standard deviation 2, proving (13). For Yat, fourth spherical moments and vanishing odd moments give the exact identity

$$\mathrm{Cov}(\mathbf{Q}) = \mathcal{L}_b(\mathbf{C}_d) - \frac{2}{d+2}\left(\mathbf{1}\mathbf{1}^\top + 2(\mathbf{C}_d \circ \mathbf{C}_d)\right).$$

Its difference from $\mathcal{L}_b(\mathbf{C}_d)$ has operator norm at most $6m/(d+2)$, since $|C_{ij,d}| \le 1$. Combining this with (15) proves (14). Finally $M_d = O(\sqrt{d})$ and $\ell_d \sim d$, which give the stated rates. □

Corollary 5 (Geometry error and finite-dimensional spectral guarantees)**.** *In Proposition 9, let* $\mathbf{C}$ *be any* $m \times m$ *correlation matrix and set* $\Delta_d = \|\mathbf{C}_d - \mathbf{C}\|_{\mathrm{op}}$*. For either input law* $P \in \{G, S\}$*,*

$$\|d^2 \,\mathrm{Cov}(\mathbf{Y}_d) - \mathcal{L}_b(\mathbf{C})\|_{\mathrm{op}} \le e_{Y,P} := E_{Y,P} + 4(1+b^2)\Delta_d.$$

If $\gamma = \lambda_{\min}(\mathcal{L}_b(\mathbf{C})) > 0$*, then*

$$\lambda_{\min}(d^2 \,\mathrm{Cov}(\mathbf{Y}_d)) \ge \gamma - e_{Y,P};$$

in particular $e_{Y,P} < \gamma$ *certifies positive definiteness. For spherical IMQ, the corresponding error relative to* $4\mathbf{C}$ *is at most* $E_{H,S} + 4\Delta_d$*, giving the analogous lower bound* $4\lambda_{\min}(\mathbf{C}) - E_{H,S} - 4\Delta_d$*. For Gaussian IMQ and* $m \ge 2$*, eigenvalues in decreasing order satisfy*

$$\lambda_1(d^3 \,\mathrm{Cov}(\mathbf{H}_d)) \ge 2m - E_{H,G}, \qquad \lambda_2(d^3 \,\mathrm{Cov}(\mathbf{H}_d)) \le E_{H,G}.$$

No positive lower eigenvalue is asserted when the comparison matrix is singular.

Proof. Schur multiplication by a correlation matrix $\mathbf{M}$ is an operator-norm contraction: factor $M_{ij} = \langle v_i, v_j \rangle$ with $\|v_i\| = 1$ and use the isometry $Ve_i = e_i \otimes v_i$ to write $\mathbf{M} \circ \mathbf{E} = V^*(\mathbf{E} \otimes I)V$. Consequently,

$$\|\mathbf{C}_d \circ \mathbf{C}_d - \mathbf{C} \circ \mathbf{C}\|_{\mathrm{op}} = \|(\mathbf{C}_d + \mathbf{C}) \circ (\mathbf{C}_d - \mathbf{C})\|_{\mathrm{op}} \le 2\Delta_d.$$

This proves the geometry-error bound. The eigenvalue inequalities follow from the variational characterization of symmetric-matrix eigenvalues and the operator-norm errors. □

For fixed b, ε the explicit bounds also give sufficient population approximation regimes $m = o(\sqrt{d})$ for Gaussian inputs and $m = o(d)$ on the sphere, relative to $\mathbf{C}_d$. Comparison with another sequence $\mathbf{C}$ additionally requires $\Delta_d \to 0$; uniform positive definiteness requires a spectral gap that dominates the displayed error. None of these statements controls empirical covariance estimation or data-dependent selection of centers.

C.3 Uniform exterior-shell separation

We return to the deterministic far-field setting. Recall the kernel

$$k_{b,\varepsilon}(\mathbf{w},\mathbf{x}) = \frac{(\mathbf{w}^\top \mathbf{x} + b)^2}{\|\mathbf{x}-\mathbf{w}\|^2 + \varepsilon}, \qquad b \geq 0, \quad \varepsilon > 0.$$

For fixed shared (b,ε), let $\mathcal{H}_{b,\varepsilon}$ denote the RKHS of $k_{b,\varepsilon}$. For $R > 0$, write

$$B_R := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq R\}, \qquad \mathbb{S}^{d-1} := \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}.$$

The earlier trace argument gives a limiting gap for each fixed IMQ expansion. Here the coefficient and center constraints make radial decay uniform over an entire comparison class, yielding L^∞ and L^2 bounds on exterior shells. The separate polynomial-separation result appears in the next appendix.

C.3.1 Asymptotic exterior-shell separation

We complement the bounded-domain gap of Section 4.2 with asymptotic L^∞ and L^2 separations on exterior shells $A_R := \{r\mathbf{u} : R \leq r \leq 2R,\ \mathbf{u} \in \mathbb{S}^{d-1}\}$ for $R \gg W$. For $W, \Lambda > 0$, write $\mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)$ and $\mathcal{R}_{\mathrm{RBF}}(\Lambda, W, \gamma)$, $\gamma > 0$, for the bounded-variation classes $\sum_{j=1}^m a_j\, h_\varepsilon(\cdot, \mathbf{v}_j)$ and $\sum_{j=1}^m a_j\, e^{-\gamma\|\cdot - \mathbf{v}_j\|^2}$ with $\sum |a_j| \leq \Lambda$ and $\|\mathbf{v}_j\| \leq W$.

Lemma 2 (Uniform radial decay on exterior shells)**.** *Let $R > W$. For $F \in \mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)$, $\sup_{A_R} |F| \leq \Lambda/((R-W)^2 + \varepsilon)$; for $F \in \mathcal{R}_{\mathrm{RBF}}(\Lambda, W, \gamma)$, $\sup_{A_R} |F| \leq \Lambda e^{-\gamma(R-W)^2}$.*

Proof. $\|\mathbf{x} - \mathbf{v}_j\| \geq R - W$ by reverse triangle inequality, so $(\|\mathbf{x}-\mathbf{v}_j\|^2 + \varepsilon)^{-1} \leq ((R-W)^2 + \varepsilon)^{-1}$ and $e^{-\gamma\|\mathbf{x}-\mathbf{v}_j\|^2} \leq e^{-\gamma(R-W)^2}$. Sum against $|a_j|$. □

Lemma 3 (Quantitative Yat far-field approximation)**.** *Fix $\mathbf{w}_\star \in \mathbb{R}^d$, $b \geq 0$, and $\varepsilon > 0$. Let $g_\star(\mathbf{x}) = k_{b,\varepsilon}(\mathbf{w}_\star, \mathbf{x})$ and $q(\mathbf{u}) := (\mathbf{u}^\top \mathbf{w}_\star)^2$. For every $R > \|\mathbf{w}_\star\|$ and every $r \in [R, 2R]$,*

$$|g_\star(r\mathbf{u}) - q(\mathbf{u})| \leq \eta_R, \qquad \eta_R := \frac{4R\|\mathbf{w}_\star\|(b + \|\mathbf{w}_\star\|^2) + b^2 + \|\mathbf{w}_\star\|^2(\|\mathbf{w}_\star\|^2 + \varepsilon)}{(R - \|\mathbf{w}_\star\|)^2 + \varepsilon}.$$

In particular, $\eta_R = O(R^{-1})$ as $R \to \infty$.

Proof. Let $W_\star := \|\mathbf{w}_\star\|$, $a := \mathbf{u}^\top \mathbf{w}_\star$, so $|a| \leq W_\star$. Then $g_\star(r\mathbf{u}) = (ra+b)^2/(r^2 - 2ra + W_\star^2 + \varepsilon)$ and $q(\mathbf{u}) = a^2$. The numerator of $g_\star(r\mathbf{u}) - a^2$ equals $(ra+b)^2 - a^2(r^2 - 2ra + W_\star^2 + \varepsilon) = 2ra(b + a^2) + b^2 - a^2(W_\star^2 + \varepsilon)$, whose absolute value is at most $2rW_\star(b + W_\star^2) + b^2 + W_\star^2(W_\star^2 + \varepsilon) \leq 4RW_\star(b + W_\star^2) + b^2 + W_\star^2(W_\star^2 + \varepsilon)$. The denominator $\|r\mathbf{u} - \mathbf{w}_\star\|^2 + \varepsilon \geq (R - W_\star)^2 + \varepsilon$. Combining gives η_R. □

Theorem 9 (Far-field separation: L^∞ and L^2)**.** *Fix $\mathbf{w}_\star \neq \mathbf{0}$, $b \geq 0$, $\varepsilon > 0$, $g_\star = k_{b,\varepsilon}(\mathbf{w}_\star, \cdot)$, $W \geq \|\mathbf{w}_\star\|$, $\Lambda > 0$, $R > W$, and let η_R be as in Lemma 3.*

(i) L^∞-bound on the exterior shell.

$$\inf_{F \in \mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)} \|g_\star - F\|_{L^\infty(A_R)} \geq \big(\|\mathbf{w}_\star\|^2 - \eta_R - \tfrac{\Lambda}{(R-W)^2 + \varepsilon}\big)_+,$$

with the analogous bound over $\mathcal{R}_{\mathrm{RBF}}(\Lambda, W, \gamma)$ replacing the IMQ decay term by $\Lambda e^{-\gamma(R-W)^2}$.

(ii) L^2 risk lower bound under directional sampling. *Let ν be a probability distribution on $\mathbb{S}^{d-1}$ with $r \in [R, 2R]$ independent of $\mathbf{u} \sim \nu$, $q(\mathbf{u}) := (\mathbf{u}^\top \mathbf{w}_\star)^2$, and $c_\nu(\mathbf{w}_\star) := \mathbb{E}[q(\mathbf{u})^2] > 0$. Then every $F \in \mathcal{R}_{\mathrm{IMQ}}(\Lambda, W)$ satisfies*

$$\mathbb{E}\big[(g_\star(\mathbf{x}) - F(\mathbf{x}))^2\big] \geq \Big[\big(\sqrt{c_\nu(\mathbf{w}_\star)} - \eta_R - \tfrac{\Lambda}{(R-W)^2 + \varepsilon}\big)_+\Big]^2,$$

with the analogous RBF bound. A single Yat atom realizes $g_\star$ at zero error in either norm.

Proof. **(i)** Take $\mathbf{u}_\star := \mathbf{w}_\star/\|\mathbf{w}_\star\|$. For $r \in [R, 2R]$, $g_\star(r\mathbf{u}_\star) \geq \|\mathbf{w}_\star\|^2 - \eta_R$ by Lemma 3 and $|F(r\mathbf{u}_\star)| \leq \Lambda/((R-W)^2+\varepsilon)$ by Lemma 2. The triangle inequality and supremum over A_R give the IMQ bound; the RBF case is identical.

(ii) $\|g_\star - q\|_{L^2} \leq \eta_R$ (Lemma 3 transferred $L^\infty \to L^2$), so reverse triangle gives $\|g_\star\|_{L^2} \geq \sqrt{c_\nu(\mathbf{w}_\star)} - \eta_R$. $\|F\|_{L^2} \leq \Lambda/((R-W)^2+\varepsilon)$ from Lemma 2. A second application of the reverse triangle inequality, followed by taking the positive part and squaring, gives the IMQ bound. Substituting the RBF decay estimate gives its counterpart. □

Remark 3. *Both bounds require $\|\mathbf{v}_j\| \leq W$ and $\sum |a_j| \leq \Lambda$; lifting either restriction allows radial methods to place centers in the shell or use coefficients growing with the radius, and the bounds become vacuous.*

D Constructive Representation and Approximation

This appendix separates exact recovery, approximation transfer, and resource constraints. The three-atom identity gives exact IMQ recovery; its minimality requires a separate proof. We then derive the span-level consequences and compare low-width Yat approximants with coefficient-constrained IMQ expansions. The final result concerns approximation of a fixed Yat section by quadratics.

D.1 Exact IMQ recovery and three-atom optimality

The bias enters only through a quadratic numerator, so a second finite difference recovers an IMQ section exactly. To prove that two atoms cannot suffice at a nonzero target center, we must also allow competing atoms at different centers. This requires irreducibility and coprimality for $d \geq 2$, and pole matching for $d = 1$.

Theorem (Theorem 2(i) restated)**.** *Fix $\varepsilon > 0$. For every $\mathbf{w} \in \mathbb{R}^d$ and every $h > 0$, the bias-finite-difference identity* (6) *holds identically in $\mathbf{x} \in \mathbb{R}^d$. In particular, $F_{\mathrm{IMQ},\varepsilon} \subseteq F^+_{\mathbb{E},\varepsilon}$. The containment is strict.*

Proof. Exact recovery. The denominator $D_\varepsilon(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$ is independent of the bias parameter. For fixed $(\mathbf{x}, \mathbf{w})$ set $t := \mathbf{x}^\top \mathbf{w} + 2h$. Then

$$g_\varepsilon(\mathbf{x}; \mathbf{w}, 3h) = \frac{(t+h)^2}{D_\varepsilon}, \quad g_\varepsilon(\mathbf{x}; \mathbf{w}, 2h) = \frac{t^2}{D_\varepsilon}, \quad g_\varepsilon(\mathbf{x}; \mathbf{w}, h) = \frac{(t-h)^2}{D_\varepsilon}.$$

The numerator second-order difference is

$$(t+h)^2 - 2t^2 + (t-h)^2 \;=\; (t^2 + 2th + h^2) - 2t^2 + (t^2 - 2th + h^2) \;=\; 2h^2,$$

independent of t. Dividing the numerator identity by D_ε gives $g_\varepsilon(\mathbf{x}; \mathbf{w}, 3h) - 2g_\varepsilon(\mathbf{x}; \mathbf{w}, 2h) + g_\varepsilon(\mathbf{x}; \mathbf{w}, h) = 2h^2/D_\varepsilon = 2h^2 k^\varepsilon_{\mathrm{IMQ}}(\mathbf{x}, \mathbf{w})$, which is (6). Since $h > 0$ the three biases $\{h, 2h, 3h\}$ are all strictly positive, so the identity uses only atoms from $F^+_{\mathbb{E},\varepsilon}$ and hence $F_{\mathrm{IMQ},\varepsilon} \subseteq F^+_{\mathbb{E},\varepsilon}$.

Strictness. Fix $\mathbf{w} \neq 0$ and a unit vector $\mathbf{u} \in \mathbb{S}^{d-1}$ with $\mathbf{u}^\top \mathbf{w} \neq 0$. Along the ray $\mathbf{x} = r\mathbf{u}$,

$$k_{\mathbb{E}}(r\mathbf{u}, \mathbf{w}) \;=\; \frac{r^2(\mathbf{u}^\top \mathbf{w})^2}{r^2 - 2r\,\mathbf{u}^\top \mathbf{w} + \|\mathbf{w}\|^2 + \varepsilon} \;\xrightarrow[r\to\infty]{}\; (\mathbf{u}^\top \mathbf{w})^2 \;\neq\; 0,$$

while every finite linear combination $F = \sum_{i=1}^N c_i k^\varepsilon_{\mathrm{IMQ}}(\cdot, \mathbf{w}_i) \in F_{\mathrm{IMQ},\varepsilon}$ obeys

$$|F(r\mathbf{u})| \;\leq\; \sum_{i=1}^N \frac{|c_i|}{r^2 - 2r\,\mathbf{u}^\top \mathbf{w}_i + \|\mathbf{w}_i\|^2 + \varepsilon} \;=\; O(r^{-2}) \;\xrightarrow[r\to\infty]{}\; 0.$$

The two far-field limits disagree, so $k_{\mathbb{E}}(\cdot, \mathbf{w}) \notin F_{\mathrm{IMQ},\varepsilon}$ for every $\mathbf{w} \neq 0$. The quadratic extrapolation identity $k_{\mathbb{E}}(\cdot, \mathbf{w}) = 3g_\varepsilon(\cdot; \mathbf{w}, h) - 3g_\varepsilon(\cdot; \mathbf{w}, 2h) + g_\varepsilon(\cdot; \mathbf{w}, 3h)$ places the same section in $F^+_{\mathbb{E},\varepsilon}$ for every $h > 0$, proving strict containment $F_{\mathrm{IMQ},\varepsilon} \subsetneq F^+_{\mathbb{E},\varepsilon}$. □

The construction uses three atoms at the target center. The next lemma controls what can happen when a competing representation uses different centers. In the complex polynomial arguments below, squared distance means the polynomial $\sum_{j=1}^{d}(x_j - w_j)^2$, not a Hermitian squared norm.

Lemma 4 (Irreducibility and coprimality of the regularized distance)**.** *Let $d \geq 2$, $\varepsilon > 0$, and for $\mathbf{w} \in \mathbb{R}^d$ set $D_{\mathbf{w}}(\mathbf{x}) := \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$. Then* (i) *$D_{\mathbf{w}}$ is irreducible in $\mathbb{C}[\mathbf{x}_1, \ldots, \mathbf{x}_d]$;* (ii) *for $\mathbf{v} \neq \mathbf{w}$, $D_{\mathbf{v}}$ and $D_{\mathbf{w}}$ are coprime in $\mathbb{C}[\mathbf{x}]$.*

Proof. (i) Translate by $\mathbf{w}$: with $\mathbf{y} := \mathbf{x} - \mathbf{w}$, $D_{\mathbf{w}} = \|\mathbf{y}\|^2 + \varepsilon$. Suppose this factors as $f(\mathbf{y})g(\mathbf{y})$ in $\mathbb{C}[\mathbf{y}]$ with $\deg f, \deg g \geq 1$. Both factors are linear: $f = \boldsymbol{\alpha}^\top \mathbf{y} + \beta$, $g = \boldsymbol{\gamma}^\top \mathbf{y} + \delta$. Matching coefficients gives

$$\boldsymbol{\alpha}\boldsymbol{\gamma}^\top + \boldsymbol{\gamma}\boldsymbol{\alpha}^\top = 2\mathbf{I}_d, \qquad \delta\boldsymbol{\alpha} + \beta\boldsymbol{\gamma} = \mathbf{0}, \qquad \beta\delta = \varepsilon.$$

Since $\varepsilon \neq 0$, neither β nor δ vanishes, so the linear constraint gives $\boldsymbol{\gamma} = -(\delta/\beta)\boldsymbol{\alpha}$ — the two direction vectors are parallel. But then $\boldsymbol{\alpha}\boldsymbol{\gamma}^\top + \boldsymbol{\gamma}\boldsymbol{\alpha}^\top = -(2\delta/\beta)\boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ has rank at most 1, while $2\mathbf{I}_d$ has rank $d \geq 2$. Contradiction; $D_{\mathbf{w}}$ is irreducible.

(ii) If $D_{\mathbf{v}}$ and $D_{\mathbf{w}}$ shared an irreducible factor for $\mathbf{v} \neq \mathbf{w}$, since both are irreducible they would equal each other up to a scalar. Matching the quadratic parts fixes that scalar at one; comparing the degree-1 part $-2\mathbf{x}^\top\mathbf{v}$ vs. $-2\mathbf{x}^\top\mathbf{w}$ then forces $\mathbf{v} = \mathbf{w}$, contradiction. □

Proof of Theorem 2(ii)

Let $d \geq 1$, $\varepsilon > 0$. For every $\mathbf{w}_0 \in \mathbb{R}^d \setminus \{0\}$, no linear combination of at most two biased Yat atoms (at arbitrary centers and biases) can equal a nonzero scalar multiple of $k^{\varepsilon}_{\mathrm{IMQ}}(\cdot, \mathbf{w}_0)$.

Outline. The argument splits by dimension. For $d \geq 2$ we use Lemma 4 (irreducibility and coprimality of $D_{\mathbf{w}}$ in $\mathbb{C}[\mathbf{x}]$) and a case-by-case polynomial analysis. For $d = 1$, D_w factors over $\mathbb{C}$ as $(z - w - i\sqrt{\varepsilon})(z - w + i\sqrt{\varepsilon})$, so Lemma 4 fails; we instead match poles directly in the complex plane.

Proof for $d \geq 2$. Write $D_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{w}_i\|^2 + \varepsilon$ for $i = 0, 1, 2$, and $L_i(\mathbf{x}) = \mathbf{x}^\top\mathbf{w}_i + b_i$ for $i = 1, 2$.

Reduction to two nonzero atoms. We first discard trivial contributions. An atom with $c_i = 0$, or with $L_i \equiv 0$ (i.e. $\mathbf{w}_i = \mathbf{0}$ and $b_i = 0$), contributes zero; dropping it gives a one-atom or zero-atom identity. A zero-atom combination is identically $0 \neq \lambda k^{\varepsilon}_{\mathrm{IMQ}}$ for $\lambda \neq 0$. For a single non-trivial atom, clearing denominators in $c\, g_\varepsilon(\cdot; \mathbf{w}, b) = \lambda\, k^{\varepsilon}_{\mathrm{IMQ}}(\cdot, \mathbf{w}_0)$ gives $c\, L^2\, D_0 = \lambda\, D_w$ in $\mathbb{C}[\mathbf{x}]$, where $L(\mathbf{x}) = \mathbf{x}^\top\mathbf{w} + b$ and $D_w = \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon$. If $\mathbf{w} = \mathbf{w}_0$: cancelling D_0 gives $c\, L^2 = \lambda$, impossible since L is nonconstant for $\mathbf{w}_0 \neq \mathbf{0}$. If $\mathbf{w} \neq \mathbf{w}_0$: by Lemma 4, D_w is irreducible (hence prime in the UFD $\mathbb{C}[\mathbf{x}]$) and coprime to D_0; from $D_w \mid c\, L^2$ we get $D_w \mid L$, but $\deg D_w = 2 > \deg L \leq 1$ and $L \not\equiv 0$: contradiction. Thus a one-atom identity cannot hold for $\lambda \neq 0$.

Assume both atoms are non-trivial, $c_i \neq 0$, $L_i \not\equiv 0$, and for contradiction $\lambda \neq 0$. Clearing denominators gives

$$c_1 L_1^2 D_0 D_2 + c_2 L_2^2 D_0 D_1 = \lambda D_1 D_2 \quad \text{in } \mathbb{R}[\mathbf{x}]. \tag{$*$}$$

Case 1: $\mathbf{w}_1 = \mathbf{w}_2 =: \mathbf{w}$. Then $D_1 = D_2 =: D$ and $(*)$ becomes $N\, D_0 = \lambda\, D$ where $N := c_1 L_1^2 + c_2 L_2^2$. *Subcase* $\mathbf{w} = \mathbf{w}_0$*:* $D = D_0$, so $N = \lambda$. The degree-2 homogeneous part of N is $(c_1 + c_2)(\mathbf{x}^\top\mathbf{w}_0)^2$. For $d \geq 2$ and $\mathbf{w}_0 \neq \mathbf{0}$ this vanishes only if $c_1 + c_2 = 0$; the degree-1 part then forces $c_1 b_1 + c_2 b_2 = 0$, so $b_1 = b_2$, and $N = b_1^2(c_1 + c_2) = 0$, contradicting $\lambda \neq 0$. *Subcase* $\mathbf{w} \neq \mathbf{w}_0$*:* by Lemma 4, D and D_0 are distinct irreducibles in $\mathbb{C}[\mathbf{x}]$, hence coprime. From $ND_0 = \lambda D$: $D \mid ND_0$ and $\gcd(D, D_0) = 1$, so $D \mid N$. Since $\deg N \leq 2 = \deg D$, we have $N = \mu D$ for some constant μ; then $\mu D_0 = \lambda$, a degree-2 polynomial equal to a constant, forcing $\mu = \lambda = 0$: contradiction.

Case 2: $\mathbf{w}_1 = \mathbf{w}_0 \neq \mathbf{w}_2$ *(the case $\mathbf{w}_2 = \mathbf{w}_0 \neq \mathbf{w}_1$ is symmetric).* Then $D_1 = D_0$; cancelling D_0 from $(*)$ gives $c_1 L_1^2 D_2 + c_2 L_2^2 D_0 = \lambda D_2$, hence

$$(c_1 L_1^2 - \lambda)\, D_2 = -c_2 L_2^2\, D_0.$$

Since $\mathbf{w}_0 \neq \mathbf{w}_2$, by Lemma 4 D_0 and D_2 are coprime irreducibles in $\mathbb{C}[\mathbf{x}]$, so $D_0 \mid (c_1 L_1^2 - \lambda)$. Both have degree $\leq 2 = \deg D_0$, so $c_1 L_1^2 - \lambda = \mu D_0$ for some $\mu \in \mathbb{C}$. Comparing degree-2 homogeneous parts: $c_1(\mathbf{x}^\top \mathbf{w}_0)^2 = \mu \|\mathbf{x}\|^2$. For $d \geq 2$ and $\mathbf{w}_0 \neq \mathbf{0}$, the polynomials $(\mathbf{x}^\top \mathbf{w}_0)^2$ and $\|\mathbf{x}\|^2$ are linearly independent in $\mathbb{C}[\mathbf{x}]$, witnessed by two real evaluations:

$\mathbf{x}$	$(\mathbf{x}^\top \mathbf{w}_0)^2$	$\|\mathbf{x}\|^2$
$\mathbf{w}_0$	$\|\mathbf{w}_0\|^4$	$\|\mathbf{w}_0\|^2$
$\mathbf{x} \neq 0,\ \mathbf{x} \perp \mathbf{w}_0$	0	$\|\mathbf{x}\|^2 > 0$

The ratios are respectively $\|\mathbf{w}_0\|^2$ and zero, so the polynomials are not proportional. The second vector exists because $d \geq 2$; this is where the coefficient comparison uses the dimensional hypothesis. At $d = 1$ these quadratic forms are proportional, so this coefficient comparison requires $d \geq 2$; all one-dimensional center configurations are treated separately below. Hence $\mu = c_1 = 0$, giving $\lambda = 0$: contradiction.

Case 3: $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2$ *pairwise distinct.* By Lemma 4, each D_i is irreducible in $\mathbb{C}[\mathbf{x}]$ and distinct D_i are pairwise coprime. Since $\mathbb{C}[\mathbf{x}]$ is a unique factorization domain (UFD), irreducibles are prime and the principal ideal (D_1) is therefore prime, making $R := \mathbb{C}[\mathbf{x}]/(D_1)$ an integral domain. Reducing $(*)$ modulo (D_1): $c_1 \overline{L_1}^2\, \overline{D_0}\, \overline{D_2} = 0$. Since $\mathbf{w}_0, \mathbf{w}_2 \neq \mathbf{w}_1$, the classes $\overline{D_0}, \overline{D_2}$ are nonzero in R; since R is an integral domain and $c_1 \neq 0$, we get $\overline{L_1} = 0$, i.e. $L_1 \in (D_1)$. But $L_1 \not\equiv 0$ and $\deg L_1 \leq 1 < 2 = \deg D_1$: contradiction. By symmetry (quotient by D_2), $c_2 = 0$: contradiction. □

Proof for $d = 1$ *via complex pole matching.* At $d = 1$ the variables collapse to scalars: $w_i, b_i, w_0 \in \mathbb{R}$, and $D_w(z) = (z - w)^2 + \varepsilon$ factors over $\mathbb{C}$ as $(z - w - i\sqrt{\varepsilon})(z - w + i\sqrt{\varepsilon})$, so Lemma 4 no longer applies. Instead, assume for contradiction

$$c_1 \frac{(w_1 z + b_1)^2}{(z - w_1)^2 + \varepsilon} + c_2 \frac{(w_2 z + b_2)^2}{(z - w_2)^2 + \varepsilon} = \frac{\lambda}{(z - w_0)^2 + \varepsilon}$$

holds for all $z \in \mathbb{R}$ with $\lambda \neq 0$, $\varepsilon > 0$. The standing hypothesis $\mathbf{w}_0 \neq \mathbf{0}$ becomes $w_0 \neq 0$ in dimension one. This excludes the origin-centered exception: for $b \neq 0$, the single atom $g_\varepsilon(\cdot; 0, b)/b^2 = k^\varepsilon_{\mathrm{IMQ}}(\cdot, 0)$ already realizes that IMQ section. Clearing denominators extends the equality to an identity of rational functions over $\mathbb{C}$.

Reduction to two nonzero atoms. First discard terms with zero coefficient or identically zero numerator. If none remain, the left-hand side is zero, contradicting the nonzero right-hand side. If only one term remains, its center must equal w_0: otherwise it is analytic at the two poles of the right-hand side. With that center, cancellation of the common denominator would give $c(w_0 z + b)^2 = \lambda$, impossible since $c \neq 0$ and $w_0 \neq 0$. We may therefore assume $c_1, c_2 \neq 0$ and both numerators are nonzero polynomials.

Pole configurations. The right-hand side has exactly two simple poles, at $z = w_0 \pm i\sqrt{\varepsilon}$. The four cases below exhaust the possible coincidences among the centers: all distinct, only the two atom centers equal, exactly one atom center equal to the target center, or all three equal.

Case 1: w_0, w_1, w_2 *pairwise distinct.* The first atom contributes potential simple poles at $z = w_1 \pm i\sqrt{\varepsilon}$. These poles are not present in either the second atom (whose poles are at $w_2 \pm i\sqrt{\varepsilon} \neq w_1 \pm i\sqrt{\varepsilon}$) nor the right-hand side (whose poles are at $w_0 \pm i\sqrt{\varepsilon} \neq w_1 \pm i\sqrt{\varepsilon}$). Hence they must be removable, requiring the linear factor $w_1 z + b_1$ to vanish at $z = w_1 + i\sqrt{\varepsilon}$:

$$w_1(w_1 + i\sqrt{\varepsilon}) + b_1 = (w_1^2 + b_1) + i w_1 \sqrt{\varepsilon} = 0.$$

Since w_1, b_1, ε are real with $\varepsilon > 0$, separating real and imaginary parts forces $w_1 = 0$ and $b_1 = 0$, i.e. the first atom is identically zero. The same argument applied to atom 2 forces $w_2 = 0 = b_2$. The identity collapses to $0 = \lambda / D_{w_0}$, contradicting $\lambda \neq 0$.

Case 2: $w_1 = w_2 =: w \neq w_0$. The two atoms share denominator $D_w(z)$. The combined left-hand-side numerator $P(z) := c_1(wz + b_1)^2 + c_2(wz + b_2)^2$ has degree ≤ 2. Equality of rational functions implies equality of pole sets with multiplicity; since $w \neq w_0$, the LHS poles at $w \pm i\sqrt{\varepsilon}$ must cancel, forcing $D_w \mid P$. Since $\deg P \leq 2 = \deg D_w$, $P = C \cdot D_w$ for

some constant $C \in \mathbb{C}$. Then LHS$= C$ is constant on $\mathbb{C}$, while RHS$= \lambda/D_{w_0}$ is non-constant (since $\lambda \neq 0$). Contradiction.

Case 3: $w_1 = w_0 \neq w_2$ (and the symmetric subcase $w_2 = w_0 \neq w_1$). The atom whose center differs from w_0 has poles at the off-w_0 locations; by the Case 1 argument applied to that atom alone, the atom is identically zero. The identity reduces to one atom $c_1(w_0 z+b_1)^2/D_{w_0}(z) = \lambda/D_{w_0}(z)$, i.e. $c_1(w_0 z + b_1)^2 = \lambda$. The left side is a polynomial in z of degree exactly 2 (using $w_0 \neq 0$ and $c_1 \neq 0$ for a non-trivial atom), while the right side is a non-zero constant. Contradiction.

Case 4: $w_1 = w_2 = w_0$. LHS and RHS share denominator $D_{w_0}(z)$; equating numerators gives the polynomial identity

$$c_1(w_0 z + b_1)^2 + c_2(w_0 z + b_2)^2 = \lambda \quad \text{in } \mathbb{R}[z].$$

Matching the z^2 coefficient: $w_0^2(c_1 + c_2) = 0$, so $c_2 = -c_1$ (since $w_0 \neq 0$). Matching the z^1 coefficient: $2w_0(c_1 b_1 + c_2 b_2) = 2w_0 c_1(b_1 - b_2) = 0$, so $b_1 = b_2$ (since $c_1 \neq 0$). Substituting back into the constant term: $c_1 b_1^2 + c_2 b_2^2 = c_1 b_1^2 - c_1 b_1^2 = 0 \neq \lambda$. Contradiction.

All four cases yield a contradiction, so the three-atom lower bound holds at $d = 1$ as well. □

D.1.1 Global uniform approximation by one growing-bias atom

Lemma 5 (Global one-atom IMQ approximation)**.** *Fix $\mathbf{w} \in \mathbb{R}^d$ and $\varepsilon > 0$, and put $W = \|\mathbf{w}\|$. For every $B > 0$,*

$$\left\| B^{-2} g_\varepsilon(\cdot; \mathbf{w}, B) - h_\varepsilon(\cdot, \mathbf{w}) \right\|_{L^\infty(\mathbb{R}^d)} \leq \frac{W/\sqrt{\varepsilon} + 2W^2/\varepsilon}{B} + \frac{W^2 + W^4/\varepsilon}{B^2}.$$

In particular, the approximation is uniform on all of $\mathbb{R}^d$ as $B \to \infty$.

Proof. Put $\mathbf{y} = \mathbf{x} - \mathbf{w}$ and $D = \|\mathbf{y}\|^2 + \varepsilon$. Expanding the numerator gives

$$B^{-2} g_\varepsilon(\mathbf{x}; \mathbf{w}, B) - h_\varepsilon(\mathbf{x}, \mathbf{w}) = \frac{2\mathbf{w}^\top \mathbf{x}}{BD} + \frac{(\mathbf{w}^\top \mathbf{x})^2}{B^2 D}.$$

Since $\mathbf{w}^\top \mathbf{x} = \mathbf{w}^\top \mathbf{y} + W^2$ and $t/(t^2 + \varepsilon) \leq 1/(2\sqrt{\varepsilon})$ for $t \geq 0$,

$$\frac{|\mathbf{w}^\top \mathbf{x}|}{D} \leq \frac{W}{2\sqrt{\varepsilon}} + \frac{W^2}{\varepsilon}.$$

Weighted Cauchy–Schwarz gives

$$(\mathbf{w}^\top \mathbf{y} + W^2)^2 \leq \left(W^2 + \frac{W^4}{\varepsilon} \right) (\|\mathbf{y}\|^2 + \varepsilon).$$

Combining these bounds and taking the supremum proves the claim. At $\mathbf{w} = \mathbf{0}$, the error is zero for every $B > 0$. □

For a compact $\mathcal{X}$, the simpler bound $(2M/B + M^2/B^2)/\varepsilon$, with $M = \sup_{\mathbf{x} \in \mathcal{X}} |\mathbf{x}^\top \mathbf{w}|$, also follows directly from the numerator expansion. Compactness is not needed for convergence. For $\mathbf{w} \neq \mathbf{0}$, the scaled atom has trace $(\mathbf{u}^\top \mathbf{w})^2/B^2$, so it is not exactly an IMQ section at any finite B. Moreover, B changes the kernel itself: the uniform limit does not assert convergence in a fixed RKHS norm.

D.2 Variable-bias span and approximation consequences

We record the span-level consequences separately for later use. Quadratic extrapolation recovers an unbiased section from positive biases; the directional trace witnesses strict containment. Exact IMQ recovery then transfers density and any supplied width-dependent approximation rate, without error from the conversion itself.

Lemma 6 (The unbiased section lies in the positive-bias span). *Fix $\mathbf{w} \in \mathbb{R}^d$ and $\varepsilon > 0$. The map $b \mapsto g_\varepsilon(\cdot; \mathbf{w}, b)$ is quadratic in b, so for every $h > 0$*

$$g_\varepsilon(\cdot; \mathbf{w}, 0) \;=\; 3\, g_\varepsilon(\cdot; \mathbf{w}, h) \;-\; 3\, g_\varepsilon(\cdot; \mathbf{w}, 2h) \;+\; g_\varepsilon(\cdot; \mathbf{w}, 3h), \tag{16}$$

where every atom on the right has bias in $\{h, 2h, 3h\} \subset (0, \infty)$. In particular $g_\varepsilon(\cdot; \mathbf{w}, 0) \in F^+_{\mathtt{E},\varepsilon}$.

Proof of Lemma 6. D_ε is independent of b, so $g_\varepsilon(\cdot; \mathbf{w}, b) D_\varepsilon = (\mathbf{x}^\top \mathbf{w} + b)^2$ is a univariate quadratic in b. Quadratic interpolation at $\{h, 2h, 3h\}$ and evaluation at $b = 0$ give coefficients $(3, -3, 1)$, since $f(0) = 3f(h) - 3f(2h) + f(3h)$ for every quadratic f. Dividing by D_ε gives (16). □

Corollary 6 (Strict span containment). *$F_{\mathrm{IMQ},\varepsilon} \subsetneq F^+_{\mathtt{E},\varepsilon}$ for every $\varepsilon > 0$. Concretely, for every $\mathbf{w} \neq \mathbf{0}$ the unbiased section $g_\varepsilon(\cdot; \mathbf{w}, 0)$ lies in $F^+_{\mathtt{E},\varepsilon}$ (Lemma 6) but not in $F_{\mathrm{IMQ},\varepsilon}$. This is the span-level companion to the RKHS-level strict containment of Corollary 1: the same separation holds at the level of finite linear spans of atoms, witnessed in both settings by the directional far-field trace. Loewner domination establishes RKHS inclusion, while the trace establishes its strictness.*

Proof of Corollary 6. The inclusion and the nonzero-center witness are proved in Theorem 2(i), Appendix D.1. That proof compares the nonzero quadratic trace with the $O(r^{-2})$ decay of every finite IMQ expansion; Lemma 6 supplies the positive-bias representation of the unbiased witness. □

Corollary 7 (Family universality, algebraic route). *For every compact $\mathcal{X} \subset \mathbb{R}^d$, both $F^+_{\mathtt{E},\varepsilon}$ at any single $\varepsilon > 0$ and the broader families $F^+_{\mathtt{E}} = \mathrm{span}\{k_{b,\varepsilon}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d, b > 0, \varepsilon > 0\}$ and $F^{\geq 0}_{\mathtt{E}} = \mathrm{span}\{k_{b,\varepsilon}(\cdot, \mathbf{w}) : \mathbf{w} \in \mathbb{R}^d, b \geq 0, \varepsilon > 0\}$ are dense in $C(\mathcal{X})$ in the uniform norm. This algebraic argument proves density of the variable-bias families; Proposition 3 establishes the stronger fixed-positive-bias universality by Loewner-IMQ domination.*

Proof of Corollary 7. $F_{\mathrm{IMQ},\varepsilon}$ is dense in $C(\mathcal{X})$ by Micchelli et al. (2006, Thm. 17). By Theorem 2, $F_{\mathrm{IMQ},\varepsilon} \subseteq F^+_{\mathtt{E},\varepsilon} \subseteq F^+_{\mathtt{E}} \subseteq F^{\geq 0}_{\mathtt{E}}$, so each superset is dense. □

The next statement transfers an existing IMQ approximation rate; it does not derive a new rate for the target class. The conversion preserves the represented function exactly and increases width by at most a factor of three. Here $|F|$ denotes the smallest number of atoms representing F in the specified family.

Corollary 8 (Approximation rate transfer at factor-3 width overhead). *Fix $\varepsilon > 0$ and $h > 0$, and let $\mathcal{X} \subset \mathbb{R}^d$ be compact. For every finite IMQ approximant $F = \sum_{j=1}^m a_j\, k^\varepsilon_{\mathrm{IMQ}}(\cdot, \mathbf{w}_j) \in F_{\mathrm{IMQ},\varepsilon}$ of m atoms, there exists a Yat approximant $G \in F^+_{\mathtt{E},\varepsilon}$ of at most $3m$ atoms (positive biases drawn from $\{h, 2h, 3h\}$) with $G(\mathbf{x}) = F(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$. In particular, any quantitative IMQ approximation rate $\inf_{F \in F_{\mathrm{IMQ},\varepsilon}, |F| \leq m} \|f - F\|_{L^\infty(\mathcal{X})} \leq \phi(m)$ transfers to a Yat rate $\inf_{G \in F^+_{\mathtt{E},\varepsilon}, |G| \leq 3m} \|f - G\|_{L^\infty(\mathcal{X})} \leq \phi(m)$. Equivalently, for every Yat budget $M \geq 3$, the transferred rate is $\phi(\lfloor M/3 \rfloor)$.*

Proof of Corollary 8. By Theorem 2, every IMQ atom equals $(2h^2)^{-1}$ times a positive-bias second-order finite difference of three Yat atoms, identically in $\mathbf{x}$. Substituting term by term in F produces a $3m$-atom Yat expansion G with $G \equiv F$ pointwise; the quantitative rate transfer is then immediate. □

D.3 Quadratic approximation and IMQ coefficient budgets

The exterior-shell separation (Appendix C.3.1) and Theorem 3 capture complementary limitations of finite radial expansions. We now prove the bounded-domain coefficient-budget theorem and extend it to mixed quadratic–IMQ targets. The obstruction places no restriction on IMQ width or center locations.

The dimension restrictions reflect two averaging arguments. Part (ii) of Theorem 3 averages over an eigenvector-orthogonal sphere and requires $d \geq 5$; part (iii) uses the full sphere and applies for $d \geq 4$.

Setup. For $\Lambda > 0$, define the coefficient-variation-bounded IMQ class

$$\mathcal{V}_{\mathrm{IMQ}}(\Lambda) \;:=\; \Big\{ \textstyle\sum_j a_j\, h_\varepsilon(\cdot, \mathbf{v}_j) \;:\; \sum |a_j| \leq \Lambda,\; \mathbf{v}_j \in \mathbb{R}^d \Big\}.$$

The atom count of an element is the smallest such expansion length.

Yat construction: move the center outward. The first two results keep the input domain fixed and increase the center norms. The width stays fixed as the tolerance decreases, but the chosen center norms grow. For fixed d, R, ε, the next bound uses center scale $O(\delta^{-1})$ as $\delta \downarrow 0$.

Lemma 7 (Single-atom Yat approximation of a rank-one quadratic)**.** *Let $\varepsilon > 0$, $\mathbf{w}_\star \in \mathbb{S}^{d-1}$, $\delta > 0$, $R > 0$, and set*

$$\alpha_0(d, R, \varepsilon, \delta) \;:=\; \max\Big\{ 8R\sqrt{d},\; 2\sqrt{dR^2 + \varepsilon},\; \tfrac{8R^3 d^{3/2}}{\delta},\; 2\sqrt{\tfrac{dR^2(dR^2+\varepsilon)}{\delta}} \Big\}.$$

For every $\alpha \geq \alpha_0$,

$$\sup_{\mathbf{x} \in [-R,R]^d} \big|\, k_{\mathbb{E}}(\alpha \mathbf{w}_\star, \mathbf{x}) - (\mathbf{w}_\star^\top \mathbf{x})^2 \,\big| \;\leq\; \delta.$$

Proof. The proof first keeps the normalized denominator away from zero, then bounds the approximation error. With $\mathbf{x} \in [-R, R]^d$, $\|\mathbf{x}\| \leq R\sqrt{d}$. Compute

$$k_{\mathbb{E}}(\alpha \mathbf{w}_\star, \mathbf{x}) \;=\; \frac{(\mathbf{w}_\star^\top \mathbf{x})^2}{1 + \eta(\mathbf{x}, \alpha)}, \qquad \eta(\mathbf{x}, \alpha) \;:=\; -\frac{2\, \mathbf{w}_\star^\top \mathbf{x}}{\alpha} + \frac{\|\mathbf{x}\|^2 + \varepsilon}{\alpha^2}.$$

On $[-R, R]^d$, $|\mathbf{w}_\star^\top \mathbf{x}| \leq R\sqrt{d}$, so

$$|\eta| \leq \frac{2R\sqrt{d}}{\alpha} + \frac{dR^2 + \varepsilon}{\alpha^2}.$$

The first two terms in the definition of α_0 make these two summands at most $1/4$ each, hence $|\eta| \leq 1/2$ for every $\delta > 0$. Therefore $|1/(1+\eta) - 1| \leq 2|\eta|$. The last two terms in α_0 give

$$dR^2 \frac{4R\sqrt{d}}{\alpha} \leq \frac{\delta}{2}, \qquad dR^2 \frac{2(dR^2 + \varepsilon)}{\alpha^2} \leq \frac{\delta}{2}.$$

Combining these estimates gives

$$\big| k_{\mathbb{E}}(\alpha \mathbf{w}_\star, \mathbf{x}) - (\mathbf{w}_\star^\top \mathbf{x})^2 \big| \leq dR^2 \left(\frac{4R\sqrt{d}}{\alpha} + \frac{2(dR^2 + \varepsilon)}{\alpha^2} \right) \leq \delta.$$

□

Proposition 10 (Spectral Yat approximation of any symmetric quadratic form)**.** *Fix $R > 0$ and $\varepsilon > 0$. Let $\mathbf{A} \in \mathrm{Sym}(d)$ have rank r with spectral decomposition $\mathbf{A} = \sum_{k=1}^r \lambda_k \mathbf{e}_k \mathbf{e}_k^\top$. For every $\delta > 0$, set $\delta_k := \delta/(r|\lambda_k|)$ and choose $\alpha_k \geq \alpha_0(d, R, \varepsilon, \delta_k)$ as in Lemma 7. Then*

$$G(\mathbf{x}) \;:=\; \sum_{k=1}^r \lambda_k\, k_{\mathbb{E}}(\alpha_k \mathbf{e}_k, \mathbf{x})$$

is an unbiased Yat expansion with at most r atoms satisfying $\sup_{\mathbf{x} \in [-R,R]^d} |G(\mathbf{x}) - \mathbf{x}^\top \mathbf{A} \mathbf{x}| \leq \delta$. For $\mathbf{A} = 0$, take the empty expansion $G = 0$.

Proof. $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{k=1}^{r} \lambda_k (\mathbf{e}_k^\top \mathbf{x})^2$. Lemma 7 applied to each unit eigenvector $\mathbf{e}_k$ with tolerance δ_k gives $|k_{\mathbb{E}}(\alpha_k \mathbf{e}_k, \mathbf{x}) - (\mathbf{e}_k^\top \mathbf{x})^2| \le \delta/(r|\lambda_k|)$. Multiplying by $|\lambda_k|$ and summing gives the displayed bound. □

IMQ obstruction: average over spheres. A uniform bound on the spherical average of each IMQ atom controls every expansion with bounded total coefficient variation, regardless of its width or center locations.

Lemma 8 (Spherical average of an IMQ section)**.** *Let $k \ge 4$, let $a > 0$ and $\tau > 0$, and let $\mathbf{U}$ be uniform on the radius-a sphere in $\mathbb{R}^k$. Then, for every $\mathbf{v} \in \mathbb{R}^k$,*

$$\mathbb{E}\frac{1}{\|\mathbf{U} - \mathbf{v}\|^2 + \tau} \le \frac{1}{a^2 + \tau}.$$

Proof. For $f_\tau(\mathbf{y}) = (\|\mathbf{y}\|^2 + \tau)^{-1}$, direct differentiation gives

$$\Delta f_\tau(\mathbf{y}) = \frac{(8 - 2k)\|\mathbf{y}\|^2 - 2k\tau}{(\|\mathbf{y}\|^2 + \tau)^3} \le 0.$$

The function $J(\mathbf{v}) = \mathbb{E} f_\tau(\mathbf{U} - \mathbf{v})$ is smooth, radial, and superharmonic by differentiation under the integral. Writing $J(\mathbf{v}) = j(r)$, smooth radial symmetry and $(r^{k-1} j'(r))' = r^{k-1} \Delta J(r) \le 0$ imply $j'(r) \le 0$. Therefore $J(\mathbf{v}) \le J(\mathbf{0}) = (a^2 + \tau)^{-1}$. □

We use this averaging bound in two ways. Slices perpendicular to a top eigenvector retain a lower bound from the largest eigenvalue; averaging on the full sphere instead captures the trace.

Proposition 11 (IMQ coefficient-budget obstruction for a PSD quadratic target)**.** *Let $d \ge 5$, let $\mathbf{A} \in \mathrm{Sym}(d)$ be PSD with $\lambda_{\max} := \lambda_{\max}(\mathbf{A}) > 0$, and fix $R, \varepsilon > 0$ and $\delta \ge 0$. If $F \in \mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ satisfies $\sup_{[-R,R]^d} |F - \mathbf{x}^\top \mathbf{A}\mathbf{x}| \le \delta$, then*

$$\Lambda \ge \sup_{0 \le t \le 1} \big(\lambda_{\max} R^2 t - \delta\big)_+ \big(R^2(1 - t) + \varepsilon\big).$$

Proof. Let $\mathbf{e}_1$ be a unit eigenvector for $\lambda_{\max}$. Fix $0 \le t < 1$ and average over the sphere

$$\mathbf{x} = R\sqrt{t}\,\mathbf{e}_1 + \mathbf{u}, \qquad \mathbf{u} \perp \mathbf{e}_1, \qquad \|\mathbf{u}\| = R\sqrt{1 - t}.$$

Let μ_t be normalized uniform surface measure on this sphere, which lies in the radius-R ball and hence in the cube. Positivity of $\mathbf{A}$ gives $\mathbf{x}^\top \mathbf{A}\mathbf{x} \ge \lambda_{\max} R^2 t$, so the spherical average of F is at least $q_t := \lambda_{\max} R^2 t - \delta$. Decompose an arbitrary center as $\mathbf{v} = s\mathbf{e}_1 + \mathbf{v}_\perp$. Lemma 8, with ambient sphere dimension $d - 1 \ge 4$, $a = R\sqrt{1 - t}$, and $\tau = \varepsilon + (s - R\sqrt{t})^2$, yields

$$\int h_\varepsilon(\mathbf{x}, \mathbf{v})\, d\mu_t(\mathbf{x}) \le \frac{1}{R^2(1 - t) + \varepsilon}.$$

Thus, whenever $q_t > 0$,

$$q_t \le \int F\, d\mu_t \le \sum_j |a_j| \int h_\varepsilon(\mathbf{x}, \mathbf{v}_j)\, d\mu_t \le \frac{\Lambda}{R^2(1 - t) + \varepsilon}.$$

The same endpoint inequality at $t = 1$ follows directly from $h_\varepsilon \le \varepsilon^{-1}$. Taking the supremum over t proves the claim. □

Proposition 12 (Trace-sensitive IMQ budget in dimension four and above)**.** *Let $d \ge 4$, $R, \varepsilon > 0$, $\delta \ge 0$, and let $\mathbf{A} \succeq 0$. If $F \in \mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ approximates $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ on $[-R, R]^d$ within δ, then*

$$\Lambda \ge \left(\frac{R^2 \operatorname{tr} \mathbf{A}}{d} - \delta\right)_+ (R^2 + \varepsilon).$$

Proof. Let $\mathbf{U}$ be uniform on the radius-R sphere in $\mathbb{R}^d$, contained in the cube. Rotational symmetry and $\|\mathbf{U}\|^2 = R^2$ give $\mathbb{E}[\mathbf{U}\mathbf{U}^\top] = (R^2/d)\mathbf{I}_d$, hence $\mathbb{E}F(\mathbf{U}) \ge R^2 \operatorname{tr} \mathbf{A}/d - \delta$. Lemma 8, now with $k = d$, $a = R$, and $\tau = \varepsilon$, gives $\mathbb{E}h_\varepsilon(\mathbf{U}, \mathbf{v}) \le (R^2 + \varepsilon)^{-1}$ for every center. Summing against absolute coefficients yields $|\mathbb{E}F(\mathbf{U})| \le \Lambda/(R^2 + \varepsilon)$. Taking the positive part proves the claim. □

For the isotropic target $\mathbf{A} = \lambda \mathbf{I}_d$ this gives $(\lambda R^2 - \delta)_+(R^2 + \varepsilon)$, whereas for a rank-one target the normalized trace is only $\lambda_{\max}/d$, and the target's spherical average is $R^2\lambda_{\max}/d$. Thus neither the trace-sensitive bound nor the eigenvector-slice bound uniformly replaces the other.

Proof of Theorem 3. Part (i) is Proposition 10. Part (ii) is Proposition 11; setting $t = 1/4$ gives the displayed special case. Part (iii) is Proposition 12. None of these arguments constrains the IMQ center locations. □

Scope of the dimension restriction. The eigenvector-slice proof applies the spherical-average lemma in $\mathbf{e}_1^\perp \cong \mathbb{R}^{d-1}$. For $k = d - 1 \geq 4$, the numerator $(8 - 2k)\|\mathbf{y}\|^2 - 2k\tau$ in Δf_τ is nonpositive everywhere, which supplies the superharmonicity used in Lemma 8. For $1 \leq k < 4$ this sign argument fails at sufficiently large radius; for $d = 1$ the positive-dimensional spherical slicing argument is unavailable. The proof therefore establishes the stated obstruction for $d \geq 5$; it does not establish failure of analogous bounds below that threshold. Proposition 12 supplies a different bound for $d = 4$; neither averaging proof establishes a counterpart in $d \leq 3$.

Explicit budget and atom normalization. The supremum in (7) is explicit. With $L = \lambda_{\max}R^2 > 0$, $S = R^2 > 0$, and any $\delta \geq 0$,

$$\sup_{0\leq t\leq 1} (Lt - \delta)_+(S(1-t) + \varepsilon) = \begin{cases} \dfrac{[L(S+\varepsilon) - \delta S]^2}{4LS}, & \varepsilon/S + \delta/L \leq 1, \\ (L-\delta)_+\varepsilon, & \varepsilon/S + \delta/L > 1. \end{cases}$$

For $0 \leq \delta < L$, maximize the concave quadratic on $[\delta/L, 1]$: its stationary point is $t_* = (1 + \varepsilon/S + \delta/L)/2$. It lies in this interval precisely when $\varepsilon/S + \delta/L \leq 1$; otherwise the maximum is at $t = 1$. For $\delta \geq L$, the objective is identically zero, as encoded by the positive part in the second branch. The obstruction depends only on total coefficient variation and does not restrict the number or locations of IMQ centers. For example, when $R = \lambda_{\max} = 1$, $\delta = 1/8$, and $\varepsilon = 1/64$, every accurate IMQ expansion must have $\Lambda \geq 3249/16384 \approx 0.1983$; the scalar supremum is attained at $t = 73/128$. This does not assert that an approximant attains that budget.

The normalization here is $h_\varepsilon = (\|\mathbf{x} - \mathbf{v}\|^2 + \varepsilon)^{-1}$; using atoms $\varepsilon h_\varepsilon$ divides the coefficients of the same expansion by ε. Under the original h_ε normalization, the concrete budget $\Lambda = 1/32$ admits no δ-accurate approximant to this target within $\mathcal{V}_{\mathrm{IMQ}}(\Lambda)$, at any finite width. The variation class itself is nonempty; it contains the zero function.

Mixed quadratic–IMQ targets. The upper bound combines the spectral construction with exact IMQ recovery, allocating half the error to each target component. For the lower bounds, the uniform size of the IMQ component enlarges the tolerance in the quadratic budget inequalities.

Theorem 10 (Yat-native approximation and IMQ budget obstruction)**.** *Let $f^\star = p^\star + g^\star$ on $[-R, R]^d$, where $p^\star(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ for a PSD $\mathbf{A} \in \mathrm{Sym}(d)$ of rank r with $\lambda_{\max}(\mathbf{A}) > 0$, and $g^\star \in \mathcal{H}^\varepsilon_{\mathrm{IMQ}}$ satisfies $\|g^\star\|_{L^\infty([-R,R]^d)} \leq M$. Assume there is a nonincreasing function $\phi : \mathbb{N} \to [0, \infty)$ with $\phi(m) \downarrow 0$ such that for every $m \in \mathbb{N}$ there exists $\bar{F}_m \in F_{\mathrm{IMQ},\varepsilon}$ using at most m atoms and satisfying*

$$\|g^\star - \bar{F}_m\|_{L^\infty([-R,R]^d)} \leq \phi(m).$$

Fix $\delta > 0$ with $\delta + M < \lambda_{\max}R^2/4$, and define the finite integer $m_\delta := \min\{m \in \mathbb{N} : \phi(m) \leq \delta/2\}$.

(i) Yat upper bound. *There exists $G \in F^{\geq 0}_{\mathbb{E}}$ with at most $r + 3m_\delta$ atoms and $\sup_{\mathbf{x}\in[-R,R]^d} |f^\star(\mathbf{x}) - G(\mathbf{x})| \leq \delta$.*

(ii) IMQ budget obstruction. *Suppose $d \geq 5$. Every $F \in \mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ with $\sup_{[-R,R]^d} |f^\star - F| \leq \delta$ necessarily satisfies*

$$\Lambda \geq \sup_{0\leq t\leq 1} \big(\lambda_{\max}R^2 t - (\delta + M)\big)_+ \big(R^2(1-t) + \varepsilon\big).$$

In particular, at $t = 1/4$,

$$\Lambda \geq \left(\frac{\lambda_{\max} R^2}{4} - (\delta + M)\right)\left(\frac{3R^2}{4} + \varepsilon\right).$$

(iii) Trace-sensitive obstruction. *For* $d \geq 4$, *every such accurate* $F \in \mathcal{V}_{\mathrm{IMQ}}(\Lambda)$ *also satisfies*

$$\Lambda \geq \left(\frac{R^2 \operatorname{tr} \mathbf{A}}{d} - (\delta + M)\right)_+ (R^2 + \varepsilon).$$

Proof. **(i)** Apply Proposition 10 to $\mathbf{A}$ with tolerance $\delta/2$, producing an unbiased Yat expansion $G_1 \in F_{\mathbb{E}}^{\geq 0}$ with r atoms and $\sup|G_1 - p^\star| \leq \delta/2$. By definition of m_δ, choose $\bar{F}_{m_\delta} \in F_{\mathrm{IMQ},\varepsilon}$ with at most m_δ IMQ atoms and $\|g^\star - \bar{F}_{m_\delta}\|_{L^\infty} \leq \delta/2$. Theorem 2 converts $\bar{F}_{m_\delta}$ into a positive-bias Yat expansion $G_2 \in F^+_{\mathbb{E},\varepsilon} \subset F_{\mathbb{E}}^{\geq 0}$ with at most $3m_\delta$ atoms and $G_2 = \bar{F}_{m_\delta}$ pointwise. Set $G := G_1 + G_2$; the triangle inequality gives the displayed atom count and error bound.

(ii) Given $\sup|f^\star - F| \leq \delta$, the bound on $g^\star$ gives $\sup|p^\star - F| \leq \delta + M$. Proposition 11 applied with tolerance $\delta + M$ gives the displayed necessary budget.

(iii) Apply Proposition 12 to the same quadratic approximation with tolerance $\delta + M$. □

Remark 4 (RKHS-norm cost of the Yat upper bound)**.** *Theorem 10(i) is an* atom-count *statement, not an RKHS-norm statement. The single-atom rank-one approximation (Lemma 7) uses a center at scale* α_0. *Choosing* $\alpha = \alpha_0$, *the corresponding section has the exact squared norm*

$$\|k_{0,\varepsilon}(\alpha_0 \mathbf{w}_\star, \cdot)\|^2_{\mathcal{H}_{0,\varepsilon}} = k_{0,\varepsilon}(\alpha_0 \mathbf{w}_\star, \alpha_0 \mathbf{w}_\star) = \frac{\alpha_0^4}{\varepsilon}.$$

In the high-accuracy regime where the δ^{-1} *term dominates, this is* $\Theta(R^{12} d^6/(\varepsilon \delta^4))$. *For the rank-r expansion of Proposition 10, the eigenvectors are orthonormal. For* $j \neq k$, *the reproducing property gives*

$$\left\langle k_{0,\varepsilon}(\alpha_j \mathbf{e}_j, \cdot), k_{0,\varepsilon}(\alpha_k \mathbf{e}_k, \cdot)\right\rangle_{\mathcal{H}_{0,\varepsilon}} = \frac{(\alpha_j \alpha_k \mathbf{e}_j^\top \mathbf{e}_k)^2}{\alpha_j^2 + \alpha_k^2 + \varepsilon} = 0.$$

Consequently the constructed approximant has the exact squared norm

$$\|G\|^2_{\mathcal{H}_{0,\varepsilon}} = \frac{1}{\varepsilon} \sum_{k=1}^{r} \lambda_k^2 \alpha_k^4. \tag{17}$$

For fixed nonzero $\mathbf{A}, d, R, \varepsilon$ *and the prescribed choices* $\alpha_k = \alpha_0(d, R, \varepsilon, \delta/(r|\lambda_k|))$, *this gives* $\|G\|_{\mathcal{H}_{0,\varepsilon}} = \Theta(\delta^{-2})$ *as* $\delta \downarrow 0$. *Larger center scales can increase this norm further. The growing radius in the Rademacher bound of Theorem 12 precludes inferring a favorable generalization guarantee from this construction alone. This is a cost of the displayed approximants, not a norm lower bound for every possible approximant. The obstruction in Theorem 10 concerns total IMQ coefficient variation, not atom count, RKHS-ball radius, or generalization error.*

D.4 Polynomial separation

The preceding construction approximates quadratics by Yat sections with growing center norms. The converse question fixes one Yat section and approximates it by degree-at-most-two polynomials. Restriction to an aligned line segment gives a positive uniform approximation gap.

Theorem 11 (Polynomial approximation gap on line-containing domains)**.** *Let* $\mathbf{w}_\star \neq 0$, $b \geq 0$, $\varepsilon > 0$, *and* $g_\star(\mathbf{x}) = k_{b,\varepsilon}(\mathbf{w}_\star, \mathbf{x})$. *Let* $\mathcal{P}_2 = \{\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{a}^\top \mathbf{x} + c : \mathbf{A} = \mathbf{A}^\top\}$ *be the space of degree-at-most-two polynomials. Suppose* $X \subset \mathbb{R}^d$ *contains a nontrivial line segment* $\{t\mathbf{v} : t \in I\}$ *with* $\mathbf{v} := \mathbf{w}_\star/\|\mathbf{w}_\star\|$ *and* $I \subset \mathbb{R}$ *a compact interval with nonempty interior. Then*

$$\inf_{P \in \mathcal{P}_2} \|g_\star - P\|_{L^\infty(X)} > 0.$$

Proof. Restriction to a line. Let $\rho := \|\mathbf{w}_\star\| > 0$. Along $\mathbf{x} = t\mathbf{v}$, $h(t) := g_\star(t\mathbf{v}) = (\rho t + b)^2/((t-\rho)^2 + \varepsilon)$. Restrictions of $\mathcal{P}_2$ to this line are univariate polynomials of degree at most two.

No exact quadratic representation. Suppose $h(t) = p(t)$ for all $t \in I$ with $p \in \Pi_2$. By real-analyticity this extends globally: $(\rho t + b)^2 = p(t)((t-\rho)^2 + \varepsilon)$. If $p \neq 0$ has degree k, the right side has degree $k+2$ (since $(t-\rho)^2 + \varepsilon$ is monic quadratic), while the left has degree two, so $k = 0$, i.e. $p = C$ constant. Then $(\rho t + b)^2 = C((t-\rho)^2 + \varepsilon)$. If $C = 0$: impossible since $\rho > 0$. If $C \neq 0$: the left side has real double root $-b/\rho$ while $(t-\rho)^2 + \varepsilon$ has nonreal roots $\rho \pm i\sqrt{\varepsilon}$; contradiction.

Positive approximation gap. Hence $h \notin \Pi_2$. Since Π_2 is closed in $C(I)$, its distance from h is strictly positive. The uniform error on X is at least the error on this line segment, giving the stated gap. □

E Layer Norms, Capacity, and Composition

We first compute the exact norm of a shared-parameter layer and bound the Rademacher complexity of a fixed RKHS ball. We then distinguish sample-dependent diagonal estimates from control of learned kernel parameters. Finally, fixing a prefix gives an exact pullback RKHS with a quotient norm; this is a layer-local description, not parameter-independent capacity control across depth.

E.1 Layer-local squared Gram norm and Rademacher complexity

Proposition 13 (Closed-form squared RKHS norm)**.** *Let $b \geq 0$, $\varepsilon > 0$, and let $f(\cdot) = \sum_{j=1}^m \alpha_j\, k_{b,\varepsilon}(\mathbf{w}_j, \cdot) \in \mathcal{H}_{b,\varepsilon}$ be a finite kernel expansion with shared (b, ε). Then*

$$\|f\|^2_{\mathcal{H}_{b,\varepsilon}} \;=\; \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}, \qquad \mathbf{K}_{ij} = k_{b,\varepsilon}(\mathbf{w}_i, \mathbf{w}_j). \tag{18}$$

The identity holds for any choice of $\{\mathbf{w}_j\}$, including repeated centers; in that case $\mathbf{K}$ may be singular, and although different coefficient vectors can represent the same function f, their difference lies in $\ker \mathbf{K}$ and the quadratic form $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$ is invariant across equivalent representations.

Proof. By bilinearity and the reproducing property,

$$\begin{aligned} \|f\|^2_{\mathcal{H}_{b,\varepsilon}} \;&=\; \sum_{i,j} \alpha_i \alpha_j \left\langle k_{b,\varepsilon}(\mathbf{w}_i, \cdot),\, k_{b,\varepsilon}(\mathbf{w}_j, \cdot) \right\rangle_{\mathcal{H}_{b,\varepsilon}} \\ &=\; \sum_{i,j} \alpha_i \alpha_j\, k_{b,\varepsilon}(\mathbf{w}_i, \mathbf{w}_j) \;=\; \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}. \end{aligned}$$

□

Computational cost. A scalar-output expansion with m centers in dimension d has $O(md)$ forward cost per input; a dense c-output readout adds $O(mc)$ work after the shared kernel evaluations. Constructing the shared Gram matrix costs $O(m^2 d)$ time and $O(m^2)$ memory. Given that matrix, one scalar-output squared norm costs $O(m^2)$ work, or $O(m^2 c)$ for all c output norms. Blockwise accumulation reduces the memory needed to evaluate these norms without storing the full Gram matrix.

Singular Gram matrices and computation. For a PSD Gram matrix, $\mathbf{K}\mathbf{K}^\dagger\mathbf{K} = \mathbf{K}$ gives the equivalent expression $(\mathbf{K}\boldsymbol{\alpha})^\top \mathbf{K}^\dagger (\mathbf{K}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$. Thus coincident or near-coincident centers affect numerical conditioning, not the RKHS identity. In floating-point computation, one may deduplicate identical centers and use a PSD-aware eigendecomposition or pseudoinverse. When $b > 0$ and the retained centers are distinct, Corollary 3 guarantees a strictly positive-definite Gram matrix and hence a Cholesky factorization in exact arithmetic. Strict positive definiteness alone does not guarantee good numerical conditioning.

Removing common logit shifts. For multiclass outputs, centering the class functions removes their common component. The same Gram identity then computes each centered norm and the quotient norm of the full tuple.

Corollary 9 (Common-logit-invariant multiclass norm)**.** *Let* $f_c = \sum_{j=1}^{m} \alpha_{jc} k_{b,\varepsilon}(\mathbf{w}_j, \cdot)$ *for* $c = 1, \dots, C$*, define* $\bar{\boldsymbol{\alpha}} = C^{-1} \sum_c \boldsymbol{\alpha}_c$*, and let* $\bar{f} = C^{-1} \sum_c f_c$*. Then*

$$\|f_c - \bar{f}\|^2_{\mathcal{H}_{b,\varepsilon}} = (\boldsymbol{\alpha}_c - \bar{\boldsymbol{\alpha}})^\top \mathbf{K} (\boldsymbol{\alpha}_c - \bar{\boldsymbol{\alpha}}).$$

These C quantities are unchanged when the same RKHS function is added to every class logit. Each is the squared RKHS norm of a centered class component. Their sum is the squared quotient norm on $\mathcal{H}^C_{b,\varepsilon}$ modulo common logits:

$$\inf_{h \in \mathcal{H}_{b,\varepsilon}} \sum_{c=1}^{C} \|f_c - h\|^2_{\mathcal{H}_{b,\varepsilon}} = \sum_{c=1}^{C} \|f_c - \bar{f}\|^2_{\mathcal{H}_{b,\varepsilon}}.$$

Proof of Corollary 9. Proposition 13 applied to $f_c - \bar{f}$ gives the identity. Adding any common $h \in \mathcal{H}_{b,\varepsilon}$ changes the mean to $\bar{f} + h$, leaving every centered function unchanged. Since $\sum_c (f_c - \bar{f}) = 0$,

$$\sum_c \|f_c - h\|^2 = \sum_c \|f_c - \bar{f}\|^2 + C \|\bar{f} - h\|^2.$$

The minimum is attained at $h = \bar{f}$, which proves the quotient identity. □

Capacity of a fixed RKHS ball. The norm identity describes an individual expansion. A capacity bound instead requires a class of functions; here that class is a radius-B ball for fixed shared kernel parameters.

Theorem 12 (Single-layer Rademacher bound)**.** *Fix $b \geq 0$, $\varepsilon > 0$. Let $\mathcal{F}_{B,b,\varepsilon} = \{f \in \mathcal{H}_{b,\varepsilon} : \|f\|_{\mathcal{H}_{b,\varepsilon}} \leq B\}$ be the RKHS ball of radius $B \geq 0$, and let $\mathbf{x}_1, \dots, \mathbf{x}_n$, $n \geq 1$, be any fixed sample satisfying $\|\mathbf{x}_i\| \leq R$. With the convention $\widehat{\mathrm{Rad}}_n(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i)$, $\sigma_i \overset{\text{iid}}{\sim} \mathrm{Unif}\{\pm 1\}$,*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_{B,b,\varepsilon}) \;\leq\; \frac{B(R^2 + b)}{\sqrt{n}\,\sqrt{\varepsilon}}. \tag{19}$$

Proof. Point evaluation and the kernel diagonal. For any $f \in \mathcal{F}_{B,b,\varepsilon}$ and $\mathbf{x} \in \mathbb{R}^d$, the reproducing property of the fixed kernel $k_{b,\varepsilon}$ gives $f(\mathbf{x}) = \langle f, k_{b,\varepsilon}(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_{b,\varepsilon}}$. By Cauchy–Schwarz,

$$|f(\mathbf{x})| \;\leq\; \|f\|_{\mathcal{H}_{b,\varepsilon}} \sqrt{k_{b,\varepsilon}(\mathbf{x}, \mathbf{x})} \;\leq\; B \sqrt{k_{b,\varepsilon}(\mathbf{x}, \mathbf{x})}. \tag{20}$$

The diagonal of the biased Yat kernel is

$$k_{b,\varepsilon}(\mathbf{x}, \mathbf{x}) \;=\; \frac{(\|\mathbf{x}\|^2 + b)^2}{0 + \varepsilon} \;=\; \frac{(\|\mathbf{x}\|^2 + b)^2}{\varepsilon},$$

since $\mathbf{x}^\top \mathbf{x} + b = \|\mathbf{x}\|^2 + b$ and $\|\mathbf{x} - \mathbf{x}\|^2 + \varepsilon = \varepsilon$. For $\|\mathbf{x}\| \leq R$,

$$\sqrt{k_{b,\varepsilon}(\mathbf{x}, \mathbf{x})} \;\leq\; \frac{\|\mathbf{x}\|^2 + b}{\sqrt{\varepsilon}} \;\leq\; \frac{R^2 + b}{\sqrt{\varepsilon}}.$$

The Rademacher average. The standard RKHS Rademacher bound for a ball of radius B follows from the reproducing property and Jensen's inequality (Bartlett & Mendelson 2002, Lemma 22, with the one-sided convention $\widehat{\mathrm{Rad}}_n(\mathcal{F}) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \sigma_i f(\mathbf{x}_i)$ used in (19)). Writing $f(\mathbf{x}_i) = \langle f, k_{b,\varepsilon}(\cdot, \mathbf{x}_i) \rangle$, Hilbert-space duality gives the factor B. Jensen's inequality then gives $\mathbb{E}\|S\| \leq \sqrt{\mathbb{E}\|S\|^2}$ for $S = \sum_i \sigma_i k_{b,\varepsilon}(\cdot, \mathbf{x}_i)$:

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_{B,b,\varepsilon}) \;=\; \frac{B}{n} \mathbb{E}_{\boldsymbol{\sigma}} \Big\| \sum_i \sigma_i k_{b,\varepsilon}(\cdot, \mathbf{x}_i) \Big\|_{\mathcal{H}_{b,\varepsilon}} \;\leq\; \frac{B}{\sqrt{n}} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} k_{b,\varepsilon}(\mathbf{x}_i, \mathbf{x}_i)}.$$

The constant is exactly 1 (no symmetrization factor of 2) under this one-sided convention. Bounding the per-sample term by $(R^2+b)^2/\varepsilon$ and taking the square root gives

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_{B,b,\varepsilon}) \;\le\; B\sqrt{\frac{1}{n}\cdot\frac{(R^2+b)^2}{\varepsilon}} \;=\; \frac{B(R^2+b)}{\sqrt{n}\,\sqrt{\varepsilon}}.$$

□

Corollary (Unbiased case)**.** *At $b=0$, $\widehat{\mathrm{Rad}}_n(\mathcal{F}_{B,0,\varepsilon}) \le BR^2/(\sqrt{n}\,\sqrt{\varepsilon})$.*

A reference kernel for parameter selection. If the shared kernel parameters vary during training, the next result embeds all admissible radius-B balls into one reference ball. Its parameter limits are fixed independently of the eventual learned values.

Proposition 14 (Uniform domination over learned kernel parameters)**.** *Fix $b_{\max} \ge 0$ and $\varepsilon_{\min} > 0$. For every $0 \le b \le b_{\max}$ and $\varepsilon \ge \varepsilon_{\min}$,*

$$k_{b,\varepsilon} \preceq k_{b_{\max},\varepsilon_{\min}}.$$

Consequently, every radius-B ball in $\mathcal{H}_{b,\varepsilon}$ embeds contractively into the radius-B ball of the single reference RKHS $\mathcal{H}_{b_{\max},\varepsilon_{\min}}$.

Proof of Proposition 14. Write $p_b(\mathbf{x},\mathbf{w}) = (\mathbf{x}^\top\mathbf{w}+b)^2$ and $h_\varepsilon(\mathbf{x},\mathbf{w}) = D_\varepsilon(\mathbf{x},\mathbf{w})^{-1}$. Both $p_{b_{\max}} - p_b = 2(b_{\max}-b)\mathbf{x}^\top\mathbf{w} + (b_{\max}^2 - b^2)$ and $h_{\varepsilon_{\min}} - h_\varepsilon$ are PSD; for the latter, use the Laplace–Gaussian mixture and the nonnegative weight $e^{-t\varepsilon_{\min}} - e^{-t\varepsilon}$. Hence the Schur-product theorem gives

$$k_{b_{\max},\varepsilon_{\min}} - k_{b,\varepsilon} = (p_{b_{\max}} - p_b)h_{\varepsilon_{\min}} + p_b(h_{\varepsilon_{\min}} - h_\varepsilon) \succeq 0.$$

The contractive RKHS embedding is the standard consequence of kernel order. □

E.2 Data-dependent capacity refinement

Keeping the observed diagonal values in the preceding Rademacher calculation avoids replacing every input norm by the worst-case radius. The kernel parameters remain fixed in this empirical bound.

Corollary 10 (Data-dependent empirical bound)**.** *For fixed shared $b \ge 0$, $\varepsilon > 0$, and any finite sample $\mathbf{x}_1,\dots,\mathbf{x}_n \in \mathbb{R}^d$, the radius-$B$ ball in $\mathcal{H}_{b,\varepsilon}$ satisfies the data-dependent bound*

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_{B,b,\varepsilon}) \;\le\; \frac{B}{\sqrt{n\varepsilon}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{x}_i\|^2+b)^2}, \tag{21}$$

computable from the sample alone and at most $B(R^2+b)/(\sqrt{n}\,\sqrt{\varepsilon})$ on any sample with $\max_i \|\mathbf{x}_i\| \le R$.

Proof of Corollary 10. The proof of Theorem 12 bounds $\widehat{\mathrm{Rad}}_n$ by $(B/\sqrt{n})\sqrt{(1/n)\sum_i k_{b,\varepsilon}(\mathbf{x}_i,\mathbf{x}_i)}$ before applying the worst-case bound on the diagonal. Using the Yat-kernel diagonal $k_{b,\varepsilon}(\mathbf{x}_i,\mathbf{x}_i) = (\|\mathbf{x}_i\|^2+b)^2/\varepsilon$ directly gives (21). □

Input scale. Corollary 10 replaces the worst-case factor R^2+b by $\sqrt{n^{-1}\sum_i(\|\mathbf{x}_i\|^2+b)^2}$. For independent isotropic Gaussian inputs with fixed d and b, the law of large numbers gives its almost-sure limit $\sqrt{(d+b)^2+2d}$ as $n\to\infty$. This population factor is asymptotic to $d+b$ as $d\to\infty$ with fixed b, giving the scale $B(d+b)/\sqrt{n\varepsilon}$ in that sequential regime. This is a probabilistic consequence of the empirical corollary, not of the deterministic radius bound in Theorem 12; the Gaussian RBF and IMQ diagonals are constant.

Learned parameters and population guarantees. The Rademacher bound is layer-local: B controls the norm of one shared-(b, ε) Yat layer and R is a data-side radius on that layer's input. If (b, ε) is selected using the same sample, Theorem 12 at the final learned values is not by itself a uniform generalization bound. Proposition 14 supplies a common reference class when training enforces $b \le b_{\max}$, $\varepsilon \ge \varepsilon_{\min}$, and a common radius B: use the reference-kernel diagonal $(R^2 + b_{\max})^2/\varepsilon_{\min}$. A bound expressed in the final learned values requires separate selection, covering, stability, or sample-splitting control.

The same distinction applies to composition: freezing a prefix after fitting it on the same data does not alone turn its pullback kernel into a sample-independent population guarantee.

E.3 Pullback RKHS structure for trained prefixes

E.3.1 Prefix-conditioned pullback

Once a prefix is fixed, the next layer's kernel pulls back to the original input space. The resulting norm is the minimum ambient norm among functions that agree on the prefix image. This distinction matters when learned centers lie outside that image.

Let $X \subset \mathbb{R}^{d_0}$ be compact. For a depth-L Yat stack, write

$$\Phi_0(\mathbf{x}) = \mathbf{x}, \qquad \Phi_\ell(\mathbf{x}) = T_\ell(\Phi_{\ell-1}(\mathbf{x})),$$

where layer ℓ has coordinates $[T_\ell(\mathbf{z})]_c = \sum_{j=1}^{m_\ell} \alpha_{\ell,j,c} k_{b_\ell,\varepsilon_\ell}(\mathbf{w}_{\ell,j}, \mathbf{z})$. Write $k_\ell(\mathbf{z}, \mathbf{z}') := k_{b_\ell,\varepsilon_\ell}(\mathbf{z}, \mathbf{z}')$ and $[\mathbf{K}_\ell]_{ij} := k_\ell(\mathbf{w}_{\ell,i}, \mathbf{w}_{\ell,j})$. Define $f_{\ell,c} := [T_\ell]_c$ on the ambient layer-input space, and collect its readout coefficients in $\boldsymbol{\alpha}_{\ell,c}$.

When learned centers $\mathbf{w}_{\ell,j}$ do not lie in the prefix range $\Phi_{\ell-1}(X)$, we view each layer coordinate as the restriction to $\Phi_{\ell-1}(X)$ of a function in the global RKHS $\mathcal{H}_{k_\ell}$ on $\mathbb{R}^{d_{\ell-1}}$. Equivalently, define the extended domain $D_\ell := \Phi_{\ell-1}(X) \cup \{\mathbf{w}_{\ell,1}, \dots, \mathbf{w}_{\ell,m_\ell}\}$, apply the RKHS construction on D_ℓ, and restrict back to $\Phi_{\ell-1}(X)$; the restriction RKHS carries the quotient norm. The norm bound below holds in either view.

Theorem 13 (Prefix-conditioned pullback RKHS: PSD, quotient norm, universality)**.** *Let Z be a Hausdorff topological space, let k be a continuous PSD kernel on Z, and let $\Phi : X \to Z$ be measurable. Define $\widetilde{k}(\mathbf{x}, \mathbf{x}') := k(\Phi(\mathbf{x}), \Phi(\mathbf{x}'))$ and $\mathcal{S} := \overline{\operatorname{span}}\{k(\Phi(\mathbf{x}), \cdot) : \mathbf{x} \in X\} \subset \mathcal{H}_k$.*

(a) PSD and norm bound. *$\widetilde{k}$ is PSD on X. For every $f \in \mathcal{H}_k$, $f \circ \Phi \in \mathcal{H}_{\widetilde{k}}$ with $\|f \circ \Phi\|_{\mathcal{H}_{\widetilde{k}}} \le \|f\|_{\mathcal{H}_k}$.*

(b) Exact quotient-norm characterization.

$$\begin{aligned} \|f \circ \Phi\|_{\mathcal{H}_{\widetilde{k}}} &= \inf\{\|g\|_{\mathcal{H}_k} : g \in \mathcal{H}_k,\ g|_{\Phi(X)} = f|_{\Phi(X)}\} \\ &= \|P_{\mathcal{S}} f\|_{\mathcal{H}_k}, \end{aligned}$$

where $P_{\mathcal{S}}$ is orthogonal projection onto $\mathcal{S}$; the infimum is attained at $P_{\mathcal{S}} f$.

(c) Universality propagates through injective continuous prefixes. *If X is compact, Φ is continuous and injective, and k is universal on $\Phi(X)$, then $\widetilde{k}$ is universal on X.*

Application to layer-ℓ Yat. *Specialising to $k = k_{b_\ell,\varepsilon_\ell}$, $\Phi = \Phi_{\ell-1}$, $\widetilde{k} = \widetilde{k}_\ell$, part (a) gives the per-layer norm bound $\|[T_\ell \circ \Phi_{\ell-1}]_c\|^2_{\mathcal{H}_{\widetilde{k}_\ell}} \le \boldsymbol{\alpha}_{\ell,c}^\top \mathbf{K}_\ell \boldsymbol{\alpha}_{\ell,c}$, with equality (via (b)) iff $f_{\ell,c} \in \mathcal{S}_\ell := \overline{\operatorname{span}}\{k_\ell(\Phi_{\ell-1}(\mathbf{x}), \cdot) : \mathbf{x} \in X\}$. For $b_\ell > 0$ and continuous injective $\Phi_{\ell-1}$, part (c) together with Proposition 3 yields universality of $\widetilde{k}_\ell$ on X.*

Proof. **(a) PSD and the feature-map norm bound.** For $\{\mathbf{x}_i\} \subset X$ and $\{\beta_i\} \subset \mathbb{R}$, $\sum \beta_i \beta_j \widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = \sum \beta_i \beta_j k(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)) \ge 0$ by PSD of k.

Let $\psi(\mathbf{z}) := k(\mathbf{z}, \cdot)$ be the canonical feature map and $\widetilde{\psi}(\mathbf{x}) := \psi(\Phi(\mathbf{x}))$, so $\widetilde{k}(\mathbf{x}, \mathbf{x}') = \langle \widetilde{\psi}(\mathbf{x}), \widetilde{\psi}(\mathbf{x}') \rangle$. The reproducing property gives $(f \circ \Phi)(\mathbf{x}) = \langle f, \widetilde{\psi}(\mathbf{x}) \rangle_{\mathcal{H}_k}$, so $f \circ \Phi \in \mathcal{H}_{\widetilde{k}}$ with the displayed norm bound.

Specialising $f = f_{\ell,c} = \sum_j \alpha_{\ell,j,c} k_\ell(\mathbf{w}_{\ell,j}, \cdot)$ and using $\|f_{\ell,c}\|^2_{\mathcal{H}_{k_\ell}} = \boldsymbol{\alpha}_{\ell,c}^\top \mathbf{K}_\ell \boldsymbol{\alpha}_{\ell,c}$ delivers the per-layer Yat-Gram bound.

(b) The minimum-norm representative. Define U on the finite span of kernel sections by $U(\widetilde{k}(\mathbf{x}, \cdot)) = k(\Phi(\mathbf{x}), \cdot)$. Agreement of all kernel-section inner products makes this definition independent of the chosen finite representation and makes U an isometry on that span. It therefore extends uniquely by density to an isometric isomorphism $U : \mathcal{H}_{\widetilde{k}} \to \mathcal{S}$. Under U, $f \circ \Phi$ corresponds to $P_{\mathcal{S}} f$, so $\|f \circ \Phi\|_{\mathcal{H}_{\widetilde{k}}} = \|P_{\mathcal{S}} f\|_{\mathcal{H}_k}$. The affine set $\{g \in \mathcal{H}_k : g|_{\Phi(X)} = f|_{\Phi(X)}\}$ equals $P_{\mathcal{S}} f + \mathcal{S}^\perp$, whose unique minimum-norm element is $P_{\mathcal{S}} f$.

(c) Transfer of uniform approximation. Since X is compact, Z is Hausdorff, and Φ is continuous and injective, $\Phi : X \to \Phi(X)$ is a homeomorphism. For $g \in C(X)$, set $h := g \circ \Phi^{-1} \in C(\Phi(X))$; by universality of k, find $f \in \mathcal{H}_k$ with $\sup_{\Phi(X)} |f - h| < \eta$. Then $f \circ \Phi \in \mathcal{H}_{\widetilde{k}}$ and $\sup_X |f \circ \Phi - g| = \sup_{\Phi(X)} |f - h| < \eta$. □

Ambient norm versus pullback norm. The orthogonal decomposition in part (b) gives the exact identity

$$\|f\|^2_{\mathcal{H}_k} = \|f \circ \Phi\|^2_{\mathcal{H}_{\widetilde{k}}} + \|(I - P_{\mathcal{S}}) f\|^2_{\mathcal{H}_k}.$$

Thus the layer Gram expression is an upper bound on the squared pullback norm, with equality exactly when the ambient layer function belongs to $\mathcal{S}$. Having every center in $\Phi(X)$ is sufficient for equality. Centers outside that image do not by themselves imply strict inequality: the criterion is the nonzero orthogonal component of the resulting function.

Why injectivity is necessary. For compact $X \subset \mathbb{R}^{d_0}$, suppose $\Phi(\mathbf{x}) = \Phi(\mathbf{y})$ at distinct points. The pullback sections at $\mathbf{x}$ and $\mathbf{y}$ coincide, so the reproducing property gives $f(\mathbf{x}) = f(\mathbf{y})$ for every $f \in \mathcal{H}_{\widetilde{k}}$. For any $g \in C(X)$,

$$\inf_{f \in \mathcal{H}_{\widetilde{k}}} \|f - g\|_{\infty, X} \geq \tfrac{1}{2} |g(\mathbf{x}) - g(\mathbf{y})|.$$

Indeed, $f(\mathbf{x}) = f(\mathbf{y})$ and the triangle inequality give $|g(\mathbf{x}) - g(\mathbf{y})| \leq 2\|f - g\|_{\infty, X}$. Taking $g(\mathbf{z}) = (\mathbf{x} - \mathbf{y})^\top \mathbf{z}$ makes this lower bound positive. Thus, when the base kernel is universal on the compact prefix image, a continuous prefix gives a universal pullback if and only if it is injective. This is a property of the full pullback RKHS; a particular finite set of trained centers specifies only a finite-dimensional subspace.

F Experimental Protocols and Results

F.1 Gaussian and fixed-norm spectral diagnostic protocol

Figure 1 uses $m = 8$ orthonormal centers, dimensions $d \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$, $b = \varepsilon = 1$, and 10^5 independent inputs for each of five seeds. To sample exactly without materializing all d coordinates, the first m Gaussian coordinates are drawn explicitly and the remaining squared norm is drawn from χ^2_{d-m}. Gaussian inputs use these coordinates and their full random norm. Fixed-norm inputs divide the same Gaussian vector by its norm and multiply by $\sqrt{d}$. For the centered activation covariance $\widehat{\Sigma}$, the reported effective rank is $\operatorname{tr}(\widehat{\Sigma})^2 / \operatorname{tr}(\widehat{\Sigma}^2)$. The supplement contains the implementation and every reported mean and standard deviation.

F.2 Frozen-CLIP diagnostics and experimental record

Experimental questions. The frozen-CLIP evidence consists of a no-training scale/bandwidth sweep and a trained-head norm–accuracy association. The supplement provides the complete configurations, implementations, fitted state, and per-configuration results.

Frozen-feature scale diagnostic. For radial units behind a learned feature map, responses depend on feature-space distances relative to bandwidth; mismatched scales can lead to activation or Gram-spectrum concentration. A no-training scale diagnostic on frozen CLIP ViT-B/32 (Radford et al., 2021) ImageNet-1k (Deng et al., 2009) representations finds that at the widest tested bandwidth the median unit coefficient of variation is 6.85×10^{-4} for RBF and 1.36×10^{-3} for IMQ, versus 9.72×10^{-2} for Yat. This measures scale sensitivity on learned features, not the Gaussian or spherical asymptotic law of the theorem. The complete protocol and results follow below.

Frozen-representation scale and bandwidth diagnostic. We probe representation-scale and bandwidth sensitivity without fitting a classifier. The public CLIP features are loaded as float32 and the kernel matrices are evaluated in float64; no additional feature normalization is applied. The Yat bias is fixed at $b = 1$ throughout both sweeps. From the frozen CLIP validation representation, we select 12,000 points and 64 disjoint centers without replacement at deterministic seed 20260804. Let $\varepsilon_0 = \text{median}_{i,j} \|\mathbf{x}_i - \mathbf{w}_j\|^2 = 1.026$. We first scale both points and centers by $s \in \{0.25, 0.5, 1, 2, 4\}$ at fixed ε_0. We then fix $s = 1$ and sweep $\varepsilon/\varepsilon_0 \in \{0.01, 0.1, 1, 10, 100\}$. For each kernel, we report the median coefficient of variation across units and the participation-ratio effective rank of the sample-centered 64×64 activation covariance, using divisor $n - 1$. The RBF convention is $\exp[-\|\mathbf{x} - \mathbf{w}\|^2/(2\varepsilon)]$. A unit enters the CV median only when the absolute sample mean exceeds 10^{-12}; the count of such units is recorded to expose numerical underflow.

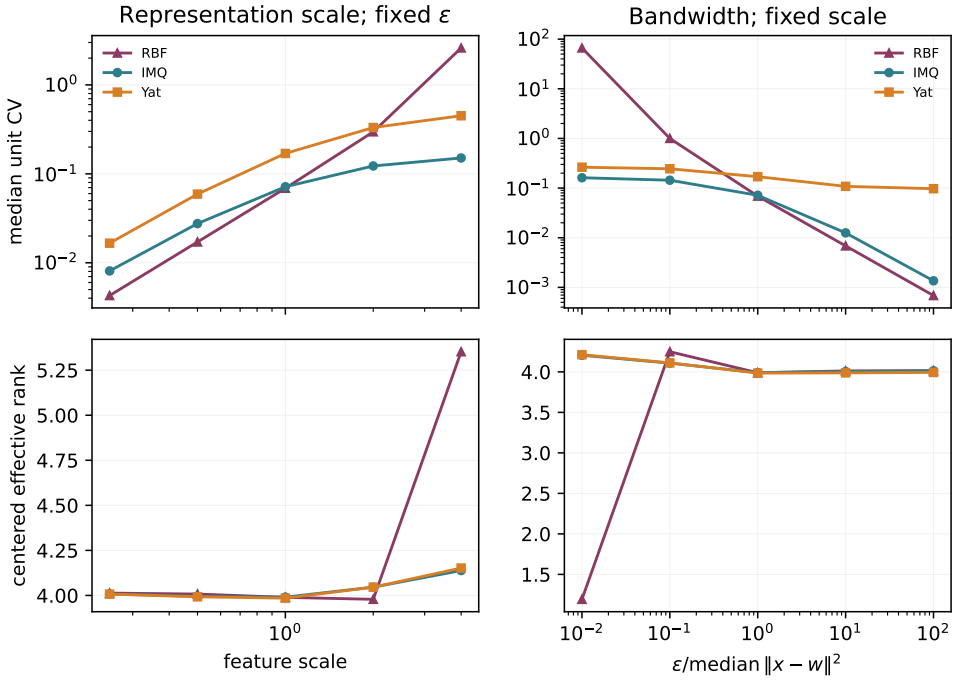

Figure 3: **Radial units show scale and bandwidth sensitivity in this frozen CLIP representation.** No classifier is trained. At $\varepsilon/\varepsilon_0 = 100$, median unit CV is 6.85×10^{-4} for RBF, 1.36×10^{-3} for IMQ, and 9.72×10^{-2} for Yat. At $\varepsilon/\varepsilon_0 = 0.01$, RBF has median CV 66.4, centered effective rank 1.19, and only 43 of 64 units with numerically nonzero mean. The effective rank of IMQ and Yat remains near 4 across most of the sweep, showing that this learned representation is far from the isotropic orthogonal-center model of Theorem 1. The measured conclusion is the relative variation supplied by the alignment numerator.

The four-order-of-magnitude bandwidth sweep displays two radial activation regimes: broad kernels concentrate their activations, while a narrow RBF can become numerically sparse. Yat is also scale-dependent: its numerator sustains more relative variation than IMQ across the recorded configurations and than RBF in the broad-bandwidth regime. Narrow RBF

kernels can instead have a larger CV alongside numerical sparsity; a larger CV alone does not imply a richer activation covariance. This diagnostic measures activation variation with centers sampled from frozen features, not classifier accuracy. The trained-head association below addresses a separate question. The supplement includes the CPU implementation and complete per-configuration statistics.

Frozen-feature norm diagnostic. A recorded width-128 Yat head on frozen CLIP ViT-B/32 ImageNet-1k representations, with fixed $b = \varepsilon = 1$, has a per-class Spearman correlation of 0.5628 between squared RKHS norm and accuracy. The correlation is 0.5571 after centering the class readouts to remove the common-logit ambiguity. These are descriptive within-run associations, not causal or generalization claims. The Adam, 30-epoch training protocol and both norm statistics are specified below. The frozen-feature diagnostics do not test the coefficient-budget obstruction.

Trained-head setup. The trained diagnostic uses frozen CLIP ViT-B/32 ImageNet-1k features ($d = 512$), a width-128 Yat head, fixed $b = \varepsilon = 1$, batch size 1024, Adam at learning rate 0.1, seed 42, and 30 epochs. Optimization uses all 1,281,167 examples in the public `train` split. The reported 73.244% accuracy and all per-class correlations use the disjoint public `validation` split of 50,000 examples. This is a within-run geometric measurement; no examples are reassigned between splits. The retained record fixes this configuration but does not contain a separate hyperparameter-selection trace. The seven-point (b, ε) stability grid repeats the same width, optimizer, epoch count, and single seed while holding each selected parameter pair fixed. Its pairs are $(0.1, 1)$, $(0.5, 1)$, $(1, 1)$, $(2, 1)$, $(1, 0.1)$, $(1, 0.5)$, and $(1, 2)$; the grid therefore includes a separate run at the reference pair $(1, 1)$.

Per-class squared RKHS norm correlation. For the recorded Yat head trained at learning rate 10^{-1}, the per-class closed-form squared RKHS norm $\boldsymbol{\alpha}_c^\top \mathbf{K} \boldsymbol{\alpha}_c$ (Proposition 13) has Spearman correlation $\rho = 0.5628$ with per-class top-1 accuracy across $n = 1000$ classes (one observation per class, 50 validation examples each). Raw class norms change when one common function is added to every logit, although softmax predictions do not. We therefore also report the common-logit-invariant statistic from Corollary 9,

$$(\boldsymbol{\alpha}_c - \bar{\boldsymbol{\alpha}})^\top \mathbf{K} (\boldsymbol{\alpha}_c - \bar{\boldsymbol{\alpha}}), \qquad \bar{\boldsymbol{\alpha}} = 1000^{-1} \sum_{c=1}^{1000} \boldsymbol{\alpha}_c.$$

Its correlation is $\rho = 0.5571$. The association therefore persists after removing the common-logit ambiguity. We treat both correlations as descriptive statistics: the 50 examples per class and the shared trained head do not define independent replicates for a conventional correlation test. Seven additional fixed-parameter, single-seed runs give raw-norm correlations in $[0.537, 0.557]$ and centered-norm correlations in $[0.530, 0.553]$ across the recorded (b, ε) grid. These descriptive associations measure fitted class geometry rather than the Rademacher bound. Both norms are computed from trained centers and readout weights without evaluation features.

The same Gram formula applies to learned-center RBF and IMQ heads. Yat contributes the diagonal $(\|\mathbf{x}\|^2 + b)^2/\varepsilon$, the directional trace, and the bias finite-difference identity; the reproducing-property calculation is shared across fixed-kernel heads.

F.3 Directional-tail benchmark: protocol and full results

The far-field formula is directly visible in a two-dimensional synthetic test. Finite IMQ expansions with 50 and 200 learned centers are fitted to one Yat atom on an annulus and evaluated farther along the aligned ray. On $r \geq 400$, their mean absolute errors are 1.008 and 1.007, respectively, against the limiting value 1.000 predicted by Corollary 4. The diagnostic illustrates the asymptotic obstruction quantitatively, but the fits retain appreciable in-annulus error; it does not test extrapolation after accurate interpolation. The construction, fitting errors, and training protocol follow below.

We illustrate Corollary 4 with a synthetic experiment in $d = 2$. The target is the single Yat atom $g_\varepsilon(\mathbf{x}; \mathbf{w}^*, b^*)$ with $\mathbf{w}^* = [1, 0]$, $b^* = 1$, $\varepsilon = 1$. Along the ray $\mathbf{u} = [1, 0]$ the alignment

factor satisfies $a = (\mathbf{u}^\top \mathbf{w}^*)^2 = 1$, so Corollary 4 predicts $|g_\varepsilon(r\mathbf{u}) - F(r\mathbf{u})| \to 1$ for every finite IMQ combination F. The experiment visualizes this far-field mechanism over a finite range of radii.

Setup. We draw $N = 4{,}000$ training points uniformly (area-measure) from the annulus $\|\mathbf{x}\| \in [50, 100]$, labels $y = g_\varepsilon(\mathbf{x})$. Two IMQ models are trained: M=50 and M=200 learned centers, shared bandwidth ε_{IMQ} trained jointly, Adam for 5,000 full-batch optimizer updates at learning rates 10^{-3} and $5{\times}10^{-4}$. The training seed is 42; centers are initialized uniformly on $[-100, 100]^2$, coefficients at zero, and bandwidth at one with parameterization $\varepsilon_{\text{IMQ}} = \exp(\eta)$. A separate 4,000-point area-uniform annulus sample uses seed 20260905 for held-out evaluation. Both models are evaluated along the ray $r\mathbf{u}$ at 500 equally spaced radii in $[0, 500]$, including both endpoints. Tail means use the 100 grid points with $r \geq 400$. The implementation is provided in the supplementary material; the MLX rerun used 4.24 and 8.08 s of training time, respectively, on the local Apple Silicon GPU.

Table 3: Mean and max pointwise error on the far tail ($r \geq 400$) for finite IMQ expansions trained to fit a single Yat atom on an annulus. The asymptotic limit predicted by Corollary 4 is $a = 1$.

Model	M	Mean $\lvert g - F\rvert$ at $r \geq 400$	Max $\lvert g - F\rvert$ at $r \geq 400$
IMQ-50	50	1.008	1.008
IMQ-200	200	1.007	1.007
Predicted ($a = 1$)	—	1.000	1.000

Results. Both fitted expansions approach error 1 on the evaluated far-field interval, consistent with Corollary 4. Increasing the width from 50 to 200 leaves the far-field limit unchanged, as the corollary predicts. The rerun also measures fit quality and parameter scale:

Model	Train MSE	Held-out MSE	Held-out MAE	ε_{IMQ}	$\max_j \|\mathbf{v}_j\|$	$\sum_j \lvert a_j\rvert$
IMQ-50	0.2290	0.2318	0.3767	4.2171	130.05	265.24
IMQ-200	0.1591	0.1856	0.3145	1.0560	132.46	457.61

These fits retain appreciable error inside the annulus. Their measured tail behavior is consistent with the predicted limit, but does not test extrapolation after accurate interpolation; that would require improved annulus fitting.

F.4 Causal-LM depth diagnostic: protocol and full results

Depth diagnostic. The theoretical results describe layer and fixed-prefix geometry; the deep experiment separately tests whether the primitive remains optimizable throughout a realistic stack. In all 12 MLP blocks of a 261M decoder-only language model, we replace the standard GELU-based hidden transformation with an $\mathbf{x} \to$ Yat $\to$ Linear block and train on 5.22B C4 tokens. The comparison matches parameter count and token budget, not learning-rate selection histories. Both Yat variants with learned output scaling obtain lower WikiText and long-range perplexity than GELU in the three-seed runs (Table 4). Training throughput is approximately 655K tokens/s for Yat and 715K for GELU on a TPU v6e-8 slice, an approximately 8% reduction. The long-range scores are historical records with unresolved dataset and checkpoint linkage, not verified named-benchmark comparisons. The runs demonstrate trainability in this setting, rather than a global capacity guarantee. The shared-bias variant retains the exact Gram norm separately at each layer; fixing the preceding network also gives the RKHS on the original input domain from Theorem 13. The full results below cover all four Yat variants, six zero-shot tasks, and the recorded compute estimates.

Setup. Using the public JAX/Flax training codebase described under compute resources and implementation below, we replace the standard GELU MLP block of a 261M decoder-

only causal language model (12 layers, n_{embd}=768, context 1024) with a $\mathbf{x} \to$ Yat $\to$ Linear block at the same parameter count. All configurations are trained for 19,920 optimizer steps on 5.22B C4 tokens (Raffel et al., 2020; Hoffmann et al., 2022), with global batch 256. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01, 500 linear-warmup steps, and cosine decay to 5% of the peak learning rate. The retained main-run configurations use lr=0.01 for GELU and lr=0.03 for Yat. We evaluate the full {pn,sb} $\times$ {$+\alpha$, +ca} Yat grid: *pn* = per-neuron bias, *sb* = bias shared within a layer, $+\alpha$ = learned output scaling, and +ca = constant output scaling.

Layer implementation and theoretical scope. The released implementation represents each hidden coordinate as

$$u_j(\mathbf{x}) = \alpha \frac{(\mathbf{w}_j^\top \mathbf{x} + b_j)^2}{\|\mathbf{x} - \mathbf{w}_j\|^2 + \varepsilon}, \qquad b_j = \text{softplus}(\beta_j), \quad \varepsilon = \text{softplus}(\eta),$$

where the sb variants constrain all β_j to one shared scalar, the pn variants learn one β_j per hidden coordinate, and η is shared within the layer. Distance evaluation and the positive transforms are performed in float32, with squared distances clamped below at zero before adding ε; the activation is then cast back to the model dtype. The raw bias is initialized at zero, so the initial positive bias is $\log 2$; η is initialized to give $\varepsilon = 10^{-3}$. The $+\alpha$ variants learn the scaling and the +ca variants use a fixed scalar. The retained declarative configuration records the constant-scaling flag but not its numerical value; those rows remain incompletely specified for historical regeneration. A zero-initialized, bias-free output projection maps the hidden coordinates back to the residual width.

The sb variants have shared positive (b, ε) and are therefore finite expansions of one Mercer kernel at each layer; Proposition 13 and the prefix-conditioned pullback interpretation apply directly. The pn variants use atom-dependent bias and are not one fixed Mercer-kernel expansion, so the shared-kernel Gram norm is not assigned to those rows. The scalar scaling is absorbed into the linear readout coefficients for the shared-kernel representation. The full transformer additionally contains attention, normalization, embeddings, and residual connections, so these RKHS statements are layer-local rather than a global capacity characterization of the complete transformer.

Reported metrics. All four variants match the GELU baseline in parameter count and training-token budget, but not in learning-rate selection history. Each row is mean $\pm$ sample standard deviation over 3 independent training seeds. The recovered evaluation notebook implements the zero-shot task scoring directly; the task-metric histories retain the scores and sample counts.

Results and scope. Across the six zero-shot tasks, the reported means and standard deviations show no uniform separation between the four Yat variants and GELU. The two $+\alpha$ variants have lower WikiText perplexity (39.0 versus 44.2) and lower long-range perplexity (101–104 versus 114) than GELU in these three-seed runs. Throughput is approximately 655K tokens/s for Yat and 715K tokens/s for GELU, measured at the training batch size

Table 4: Parameter- and token-budget-matched 261M causal LMs trained on 5.22B C4 tokens; mean $\pm$ standard deviation over three seeds. Perplexity is lower-is-better. Throughput is measured at the training batch size on a TPU v6e-8 slice. These are feasibility runs at retained settings, not matched-hyperparameter-tuning comparisons. Here pn denotes per-neuron bias, sb denotes bias shared within a layer, and $+\alpha$ denotes learned output scaling. The historical long-range dataset source is unverified; the evaluation protocol and provenance limitations are recorded below.

Primitive	WikiText PPL	Long-range PPL	Tokens/s
GELU	44.23 ± 1.44	114.08 ± 2.82	715K
Yat (pn+α)	39.04 ± 1.03	104.20 ± 3.40	655K
Yat (sb+α)	39.07 ± 0.39	101.09 ± 3.89	655K

Table 5: **261M causal-LM evaluation, 5.2B C4 tokens, mean ± sample standard deviation over 3 training seeds.** Wiki PPL scores 29,970 next tokens; long-range PPL scores 16,380 next tokens in four 4096-token chunks. The recovered evaluator uses WikiText-2 for the former and prioritizes PG-19 for the latter (Rae et al., 2020); source-selection details are given below; the actual historical long-range source is unverified. These are not matched-tuning comparisons. Perplexity is lower-is-better; zero-shot accuracy is higher-is-better.

Perplexity

Variant	Wiki PPL ↓	Long-range PPL ↓
GELU	$44.23_{\pm 1.44}$	$114.08_{\pm 2.82}$
Yat (pn+α)	$39.04_{\pm 1.03}$	$104.20_{\pm 3.40}$
Yat (sb+α)	$39.07_{\pm 0.39}$	$101.09_{\pm 3.89}$
Yat (pn+ca)	$42.18_{\pm 0.28}$	$108.62_{\pm 3.51}$
Yat (sb+ca)	$42.09_{\pm 1.96}$	$106.51_{\pm 4.08}$

Zero-shot accuracy: language and ARC tasks

Variant	LAMBADA ↑	ARC-E ↑	ARC-C ↑
GELU	$23.59_{\pm 0.87}\%$	$33.53_{\pm 1.12}\%$	$22.47_{\pm 0.90}\%$
Yat (pn+α)	$24.16_{\pm 0.43}\%$	$33.01_{\pm 0.80}\%$	$22.01_{\pm 1.67}\%$
Yat (sb+α)	$23.80_{\pm 1.20}\%$	$32.62_{\pm 0.84}\%$	$21.39_{\pm 0.81}\%$
Yat (pn+ca)	$23.79_{\pm 0.71}\%$	$33.02_{\pm 0.43}\%$	$22.07_{\pm 0.79}\%$
Yat (sb+ca)	$23.43_{\pm 0.83}\%$	$32.98_{\pm 1.47}\%$	$22.44_{\pm 0.31}\%$

Zero-shot accuracy: commonsense tasks

Variant	HellaSwag ↑	OpenBookQA ↑	PIQA ↑
GELU	$29.00_{\pm 0.05}\%$	$27.87_{\pm 1.63}\%$	$61.77_{\pm 0.65}\%$
Yat (pn+α)	$29.62_{\pm 0.52}\%$	$28.20_{\pm 2.23}\%$	$62.44_{\pm 0.56}\%$
Yat (sb+α)	$29.52_{\pm 0.18}\%$	$28.40_{\pm 0.35}\%$	$62.37_{\pm 0.74}\%$
Yat (pn+ca)	$29.20_{\pm 0.39}\%$	$27.47_{\pm 1.40}\%$	$61.26_{\pm 1.07}\%$
Yat (sb+ca)	$29.62_{\pm 0.18}\%$	$27.20_{\pm 0.20}\%$	$61.50_{\pm 0.69}\%$

on a TPU v6e-8 slice and averaged over the full run. Thus the tested Yat parameterizations can be optimized across 12 layers, with an approximately 8% throughput reduction. For the shared-parameter rows, Proposition 13 remains applicable separately at each layer.

Application questions beyond the diagnostics. Further application questions remain separate from these mathematical claims. Whether changes to individual components produce predictable effects on held-out inputs requires intervention tests; whether specialized implementations improve latency, memory, or energy at matched quality requires end-to-end systems comparisons. Neither follows from the exact layer geometry alone.

Compute resources and implementation. Each 261M variant was trained on 5.22B C4 tokens, approximately 20 training tokens per parameter. Per-variant training used bfloat16 model computation in JAX/Flax on a Google Cloud TPU v6e-8 slice (TPU Research Cloud), while numerically sensitive reductions such as logits and the loss were accumulated in float32. Compute below is reported in TPU-v6e-8 slice-hours: one hour on the full eight-chip slice counts as one slice-hour. The reported token counts and rounded throughput imply about 2.03 slice-hours per GELU run and 2.21 per Yat run, or 32.65 across the fifteen main runs. This is a throughput-based estimate. Measured elapsed times for every main run and the complete tuning campaign are needed to determine total training compute. The supplementary validation record checks the throughput arithmetic. The anonymous artifact contains the declarative launch specification; the public build provides the identity-bearing checkpoint/run map.

Aggregate reconstruction and historical provenance. The supplement includes the 15 per-run task-metric histories for this comparison. Selected complete eight-task passes provide all 40 displayed task-metric cells. We compute the mean and sample standard deviation ($n-1$ denominator) from full-precision seed metrics and round only the final

values to two decimal places. For 261M, these are the second complete pass for seed zero and the first complete pass for seeds one and two. These whole passes match the earlier table's recorded scores; the historical reason for selecting them remains undocumented. The recovered notebook resumes runs by model repository and sample limit, without including a checkpoint revision in the run identifier or model-loading call. Rerunning the models additionally requires the original training revision and environment lockfile.

Evaluation protocol. The histories record 29,970 scored tokens for Wiki PPL and 16,380 for long-range PPL, with the latter using four 4096-token chunks. The recovered notebook computes Wiki PPL from the first 30,000 tokens of WikiText-2 test, in nonoverlapping 1024-token chunks. For long-range PPL it tries PG-19 validation, then PG-19 test, then WikiText-103 test. The selected source and executed notebook revision were not recorded with the task scores, so the long-range column is identified by its logged context protocol rather than assigned a verified dataset split.

The zero-shot counts are 5153 LAMBADA examples, 2376 ARC-Easy, 1172 ARC-Challenge, 10,042 HellaSwag, 500 OpenBookQA, and 1838 PIQA.

Learning-rate sweep. The recovered 261M Yat precursor sweep tests ten peak rates, $\{10^{-4}, 3{\times}10^{-4}, 10^{-3}, 3{\times}10^{-3}, 10^{-2}, 0.03, 0.1, 0.3, 1, 3\}$. Each completed candidate uses 1000 optimizer updates, global batch 256 and context 1024: 262,144,000 token slots, with padding excluded from the loss. AdamW uses weight decay 0.01, gradient clipping at norm 1, 100 linear-warmup updates, and cosine decay toward 5% of the peak rate. Among these candidates, 0.03 has the lowest recorded final-minibatch loss (4.5621, versus 4.8306 at 0.01), consistent with the retained Yat main-run rate. The script also records an exponential loss average with coefficient 0.95; it does not encode an automatic selection rule. The GELU main-run rate is 0.01; the recovered sweep evidence does not establish its selection history or a separate sweep for each Yat variant.

Checkpoint availability. External Hugging Face repositories archive all three 261M checkpoints for every table row; an external W&B project retains the training curves and logged metrics. Identity-bearing code, model, and W&B links will be supplied in the non-anonymous version.

G Reproducibility and Public Artifacts

Appendices F.1–F.4 document the recorded protocols and results for the spectral and frozen-CLIP diagnostics, directional-tail test, and causal-LM experiments. The supplement provides executable implementations, per-configuration results, and fitted states for the retained CLIP experiments.

The causal-LM experiments use the JAX/Flax codebase FlaxChat. The training codebase is externally public; checkpoints, training curves, and logged metrics are retained in external Hugging Face and W&B projects. Identity-bearing links are omitted from the anonymous version. The public supplementary map links all 15 paper row/seed combinations to immutable retained Hugging Face checkpoint revisions and W&B training runs. This identifies the retained training checkpoint for each row/seed, but does not establish that each historical evaluation loaded that exact revision: the original evaluator did not log the loaded checkpoint revision. The release also includes a declarative launch specification for every reported configuration and per-seed evaluation histories from which the reported task metrics can be reconstructed.

Source scripts and machine-readable diagnostic records are retained in the experiment directory of the canonical source. The ICLR package includes the figures and local bibliography needed to rebuild the manuscript with XeLaTeX.

Reconstructing the reported metrics is distinct from rerunning the original evaluations. The recovered evaluation notebook did not pin checkpoint revisions; rerunning the models additionally requires the original training revision and environment lockfile. Appendix F.4

details these historical-linkage limitations, the recorded evaluation protocol, and the available execution records.